NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

MULTIPLE ADDITIVE REGRESSION TREES
– A METHODOLOGY FOR PREDICTIVE DATA MINING –
FOR FRAUD DETECTION

by

António S. Monteiro

September 2002

Thesis Advisor:	Lyn R. Whitaker
Second Reader:	Samuel E. Buttrey

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704 – 0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY ***(Leave Blank)***	**2. REPORT DATE** September 2002	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE: **Multiple Additive Regression Trees - a Methodology for predictive Data Mining - for fraud detection**	**5. FUNDING NUMBERS**
6. AUTHOR(S) *Monteiro, António Jorge Ferreira da Silva*	
7. PERFORMING ORGANIZATION NAME(S) AND ADRESS(ES) Naval Postgraduate School Monterey, CA 93943 - 5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Operation Mongoose / DFAS 400 Gigling road Seaside, CA 93955 - 6771	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES
The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.	**12b. DISTRIBUTION CODE**

13. ABSTRACT ***(maximum 200 words)***

The Defense Finance Accounting Service DFAS-Operation Mongoose (Internal Review - Seaside) is using new and innovative techniques for fraud detection. Their primary techniques for fraud detection are the data mining tools of classification trees and neural networks as well as methods for pooling the results of multiple model fits. In this thesis a new data mining methodology, Multiple Additive Regression Trees (MART) is applied to the problem of detecting potential fraudulent and suspect transactions (those with conditions needing improvement – CNI's). The new MART methodology is an automated method for pooling a "forest" of hundreds of classification trees. This study shows how MART can be applied to fraud data. In particular it shows how MART identified classes of important variables and that MART is as effective with raw input variables as it is with the categorical variables currently constructed individually by DFAS. MART is also used to explore the effects of the substantial amount of missing data in the historical fraud database. In general MART is as accurate as existing methods, requires much less effort to implement saving many man-days, handles missing values in a sensible and transparent way, and provides features such as identifying more important variables.

14. SUBJECT TERMS Fraud, Data Mining, MART, Classification Trees, Relative importance of variables, Missing values.	**15. NUMBER OF PAGES** 113
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

THIS PAGE INTENTIONALLY LEFT BLANK

MULTIPLE ADDITIVE REGRESSION TREES
– A METHODOLOGY FOR PREDICTIVE DATA MINING –
FOR FRAUD DETECTION

by

António S. Monteiro
Lieutenant Commander, Portuguese Navy
B.S., Portuguese Naval Academy, 1990
M.S. in Statistics and Information Management, UNL–ISEGI, 1995

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 2002

Author: António Jorge Ferreira da Silva Monteiro

Approved by: Lyn R. Whitaker
Thesis Advisor

Samuel E. Buttrey
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Defense Finance Accounting Service DFAS-Operation Mongoose (Internal Review - Seaside) is using new and innovative techniques for fraud detection. Their primary techniques for fraud detection are the data mining tools of classification trees and neural networks as well as methods for pooling the results of multiple model fits. In this thesis a new data mining methodology, Multiple Additive Regression Trees (MART) is applied to the problem of detecting potential fraudulent and suspect transactions (those with conditions needing improvement – CNI's). The new MART methodology is an automated method for pooling a "forest" of hundreds of classification trees. This study shows how MART can be applied to fraud data. In particular it shows how MART identified classes of important variables and that MART is as effective with raw input variables as it is with the categorical variables currently constructed individually by DFAS. MART is also used to explore the effects of the substantial amount of missing data in the historical fraud database. In general MART is as accurate as existing methods, requires much less effort to implement saving many man-days, handles missing values in a sensible and transparent way, and provides features such as identifying more important variables.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGEMENT

I would like to acknowledge some people who were fundamental in the completion of this work. To DFAS-Operation Mongoose (Internal Review - Seaside) for supporting this thesis offering the opportunity to work on a real and promising project, with professional and enthusiastic people. To Dave Riney, Randy Faulkner and LTC Chris Drews for their support and interest along this research.

To Dr. Lyn R. Whitaker for her untiring guidance, support and confidence that greatly contributed to the successful completion of this thesis, and to Dr. Samuel E. Buttrey for his support and patient assistance rendered during this study.

Most of all I would like to thanks my wife Dores Rosa for her encouragement, support and understanding that helped me through two demanding years at NPS. To my son André for showing me the important things in life! I hope from now on you will see more of me.

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Classification and Regression Trees (CART) and Neural Networks (NN) are the primary techniques used by Defense Finance and Accounting Service (DFAS), Operation Mongoose (Internal Review – Seaside), for fraud detection. In this thesis a new data mining methodology, Multiple Additive Regression Trees (MART) is applied to the problem of detecting potential fraudulent and suspect transactions (those with conditions needing improvement – CNI's). The new MART methodology is an automated method for pooling a "forest" of hundreds of classification trees. This study shows how MART can be applied to fraud data. In particular it shows how MART identifies classes of important variables and that MART is as effective with raw input variables as it is with the categorical variables currently constructed individually by DFAS. MART is also used to explore the effects of the substantial amount of missing data in the historical fraud database. In general MART is as accurate as existing methods, requires much less effort to implement saving many man-days, handles missing values in a sensible and transparent way, and provides features such as identifying more important variables. This study helps identify improvements in the ongoing process of data mining, increasing the potential of DFAS to detect and contributing to DoD's goal of combating and eliminating fraud.

This thesis shows the applicability of MART methodology in identifying fraud. The thesis also describes the process of identifying the set of variables needed for an accurate classification with the MART approach without losing significant information. This process is used to explore whether categorical variables created from original numeric variables, in a labor-intensive process, actually improve modeling with MART. A third major concern explored here, is to whether the current classification models, based on the historical Knowledge Base, are detecting the differences between fraud and nonfraud patterns or whether they are classifying using other features that differentiate the Knowledge Base from the site-specific nonfraud cases. Fourth, the missing values' pattern in specific fields of the Knowledge Base is analyzed including a report of their role in the process of classifying fraud patterns. A summary of results from these analyses follows.

The missing values analysis of the Knowledge Base study reveals the relationship and importance missing values have on fraud classification. The nonrandom pattern of missing values might contribute to difficulties in fraud prediction. In particular this research gives some insights about the way missing values contribute for increasing the odds of observing fraud. In addition, a study of imputing missing values support the way MART methodology handles the missing values problem and reveals that no fraud prediction advantage is offered by imputing values on missing valued predictors present in the actual Knowledge Base.

The identification of relative importance of variables for classifying fraud and CNI's highlights the most relevant predictors present in the Knowledge Base and CNI database. The study of MART model performance when trained on numerical versus categorical variables for predicting fraud supports the fact that data do not require being transformed, or preprocessed in any way, for MART training purposes. This fact reveals a major advantage DFAS can explore using this methodology, saving time used to convert numeric variables into categorical ones. MART models trained on sets of numeric variables performed about as well as the MART models trained on sets of categorical variables.

The comparison of performances of different models so far developed by DFAS and the MART models trained on the Knowledge Base for several different auditing sites reveal that, in general, MART performance is comparable to other models. The importance of this is that MART takes significantly less time and manpower to use than methodologies currently used by DFAS.

In general, the MART methodology is shown to be an alternative tool for improving the current process of predicting fraud and CNI's. This methodology should be seen in an integrated knowledge environment where additional information and process improvements are offered. In addition, the insight that the Knowledge Base is potentially training models to classifying and predicting patterns other than fraud contributes to arguments that the Knowledge Base repository should be updated as current fraud cases became available. This will help identify changes or mutations in fraud patterns motivated by fraud perpetrators' intelligence as well as by technology evolution or new process transactions.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. PURPOSE

The Knowledge Discovery in Databases (KDD) process is described by Mannila [17] as being an iterative process, in which data mining is seen as one integrated step associated with pattern discovery. Data mining is being used to discover patterns and relationships in data, supporting the idea of learning from data. This relatively new and rapidly changing discipline constitutes an interdisciplinary research area lying at the intersection of statistics, machine learning, data management and databases, pattern recognition, artificial intelligence, and other areas [10]. Predictive data mining is concerned with constructing statistical models from historical data in order to predict future unknown data values, and / or to help gain insights about the predictive relationships presented in the data.

This thesis will study the use of a new data mining methodology for fraud detection. Classification Trees (CART) and Neural Networks (NN) are the primary techniques used by Defense Finance and Accounting Service (DFAS) Operation Mongoose (Internal Review – Seaside), for fraud detection. A recently developed technique by Friedman [7], [8], [9], Multiple Additive Regression Trees (MART), offers an alternative approach to classification problems. Particularly attractive for detecting fraud and transactions with conditions needing improvements (CNI's) are MART's ability to handle missing data, operate with continuous as well as categorical predictor variables, and evaluate the relative importance of predictor variables. In addition, this research will benefit the sponsor, DFAS-Operation Mongoose (Internal Review – Seaside) (referred to as "DFAS" for the remainder of the thesis), by incrementing the understanding of the fraud detection process. The analysis of the existing set of transactions known to be fraudulent (the "Knowledge Base" (KB)) and those with CNI's using MART suggests new directions on the fraud detection process. This study helps identify improvements in the ongoing process of data mining, increasing the potential of the Project Mongoose in fraud detection, and contributing to DoD's goal of combating and eliminating fraud.

B. BACKGROUND

Every government agency that trades with citizens, vendors or service providers risks exposure to irregularities and fraud. The risk of losing more and more money, through fraud, increases every year. It is to be hoped that agencies can identify fraudulent activity by “mining” their existing data.

The Defense Finance and Accounting Service has incorporated data mining into their investigating and auditing processes using historical data and expert knowledge to identify fraud patterns through analytical techniques. Paying billions of dollars worth of military bills each year, DFAS is exploring the use of data mining as a way to discover billing errors and fraud out of the millions of transactions that DOD processes each year. The Defense Finance and Accounting Service’s Operation Mongoose with the cooperation of the Defense Manpower Data Center (DMDC) are undertaking the process of fraud detection.

Currently the process of classifying fraud cases is based on training data sets in which the known fraud cases came from a small historic (1989 to 1997) database called the Knowledge Base. The nonfraud cases in the training data set are chosen from the current transactions of particular sites under study. There are several issues that need to be addressed when using these training data to build models for fraud detection. The first is whether models built on the training data are really distinguishing between fraud and nonfraud, or whether they are classifying using other features that differentiate the Knowledge Base of fraud cases from the site-specific nonfraud cases. The second issue relates to significant numbers of missing values in specific fields of the Knowledge Base, and their role in classifying fraud patterns. A third issue is that of variable selection. Currently models are built based on over 57 variables. Many of these variables are redundant, have missing data or are noninformative. In particular a great deal of effort is made to categorize continuous variables before using them for data mining. These potentially extraneous variables make model fitting a long and arduous task. Anything that can be done to make the list of variables more manageable will be a great benefit to DFAS. Finally, there is the issue about how MART compares the currently used methods, classification trees (which tend to have high misclassification rates) or neural networks (which can be affected by too many redundant or noninformative predictor variables).

This work does not propose that MART take the place of the models currently used by DFAS. The Rashomon Effect, as presented by L. Breiman [1], describes perfectly the reasoning of having a multitude of different models, different predictor subsets, each one telling a different story about the same unknown reality. The problem of identifying the best model is not a concern of the present work. Indeed, presumably no model will dominate all other.

C. OPERATION MONGOOSE – DFAS

As reported by Shawn [22], historically, fraud in DoD has not received much attention. Action was taken seriously only when occasional isolated cases became highly publicized. The DoD has historically had weak internal controls and a lack of financial accountability, which are primary facilitators of fraud.

To enhance DoD’s fraud detection capability, Operation Mongoose was created in 1994 with the primary purpose of detecting fraud in retired and annuitant pay, military pay, civilian pay, transportation payments, and vendor payments. Since its inception, it has analyzed tens of millions of financial transactions to detect potential cases of error and fraud. Operation Mongoose is the first multi-agency program formed with national scope to examine possible financial fraud. In order to assist in combating fraud, many federal agencies have combined forces to form task groups and /or share information. For example, Operation Mongoose, the DoD fraud detection unit, has formed an alliance with the Defense Manpower Data Center (DMDC) and DoD Inspector General (DoDIG), the United States Secret Service, and service audit agencies. The Defense Finance and Accounting Service (DFAS) established a partnership with the Defense Criminal Investigative Service (DCIS) and the Air Force Audit Agency. In February 1998, DFAS also created a Fraud Task Force [22].

The alliance with DMDC offers Operation Mongoose access to payment and supporting data on several different computer systems from different sites. Indicators developed by subject matter experts from DFAS, DMDC and the DoDIG, when matched against transactions' information from the different DoD payment systems, help to identify anomalies in the data which can in turn to expose fraud or internal control weaknesses.

The different DoD vendor pay systems, Automated Financial Entitlement System (AFES), the Computerized Accounts Payable System (CAPS), the Integrated Accounts Payable System (IAPS) and the Standard Accounting and Reporting System-Field Level (STARS-FL) constitute the data sources. Information related to invoice, receipt and vouchers detail each transaction with information such as unit price, quantity ordered, merchandise receipt and costs, dates of receipt, freight amounts, vendor name, address and payment type, item description, voucher number, and some accounting data.

D. SCOPE OF THE THESIS

This thesis first, analyzes the applicability of MART methodology in data mining to identify fraud, and second describes the process of identifying the set of variables needed for an accurate classification with the MART approach without losing significant information. A third goal addressing a major concern, is to analyze whether the current classification models, based on the historical Knowledge Base, are detecting the differences between fraud and nonfraud patterns or whether they are classifying using other features that differentiate the Knowledge Base from the site-specific nonfraud cases. Fourth, the missing values' pattern in specific fields of the Knowledge Base is analyzed including a report of their role in the process of classifying fraud patterns. Finally MART is compared to existing DFAS models.

E. OUTLINE OF THESIS

The remainder of the thesis is organized as follows. Chapter II provides a brief description of types of fraud, Conditions Needing Improvement (CNI) and how the Knowledge Base was constructed.

Chapter III presents a detailed overview of the MART methodology focusing on features pertinent to improving the knowledge about the process of predicting fraud and CNI's, with examples.

Chapter IV presents the results of the statistical analysis of the Knowledge Base including a missing value analysis. An analysis about whether the classification models, based on the historical Knowledge Base, are detecting the differences between fraud and nonfraud patterns or whether they are classifying using other features that differentiate the Knowledge Base from the site specific nonfraud cases, is found in this chapter. In addition, the approach MART uses for handling missing values is compared to other methods.

Chapter V approaches the problem of identifying the relative important predictors for fraud and CNI prediction, based on the available datasets. A description and discussion of the most significant results are presented here. Also, related results detailing some methods that will save development time working with continuous variables are presented. The chapter concludes with an overview and comparison of the MART models performance with C5 and NN models developed by DFAS's expert data mining team.

Finally in Chapter VI this work on classification and fraud prediction is summarized. Findings of the research, and recommendations are presented for further research and study.

THIS PAGE INTENTIONALLY LEFT BLANK

II. FRAUD CLASSIFICATION

DFAS staff uses three types of data: historical fraud data, data on all current transactions and information from audits on selected transactions. They obtain current data from the DMDC, which has the ability to gather data from a variety of computer systems from different sites giving IR access to transaction information by field site and vendor pay system for a period of eighteen months. IR also uses subject matter experts from DFAS, DMDC and DODIG to develop additional indicators which are incorporated into the data.

The second major source of transaction data used by DFAS is the historical database of known fraud, the Knowledge Base. This data consists of 442 transactions involving a total of 21 fraud cases detected between 1989 and 1997. These cases cover fraud committed using fake documents (false invoices, false certification of receipts, false purchase requests and false vouchers), false employees, and altered documents (overpayment, resubmission). The following figures present the distribution (percentage of cases) of fraud type and amounts stolen described and tabled in Shawn [22].

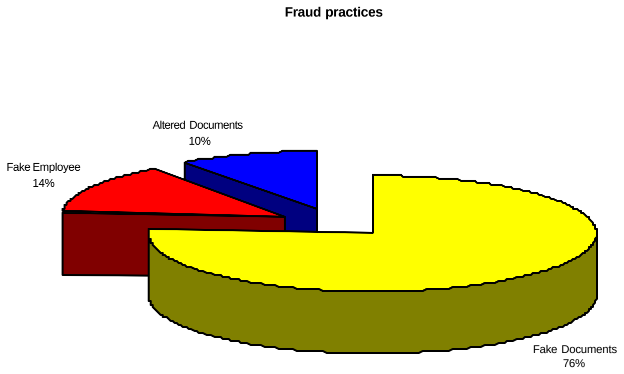


Figure 2.1 Fraud practices

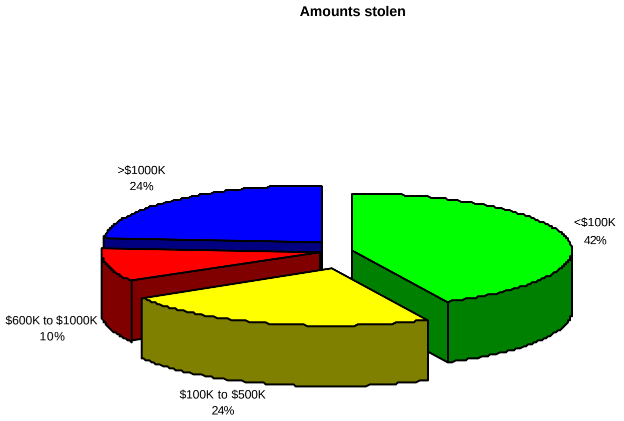


Figure 2.2 Amounts stolen

More detailed information about fraud discovery sources, management control violations, status of perpetrator and associated service (Navy, Army, Air Force and Air National Guard) are presented in Shawn [22]. A detailed description and discussion of the Knowledge Base can be found in Jenkins [12], Shawn [22].

Detailed information associated with those reported fraud cases constitutes the basis of the fraud historic repository known as Knowledge Base. Some of those fraud cases present more than one irregular transaction, adding up to a total of 442 fraud transactions.

A. CLASSIFICATION OF FRAUD IN THE KNOWLEDGE BASE

In the initial construction of the Knowledge Base, fraud was classified according to six fraudulent payment types rather than using a single fraud/nonfraud binary classification. Experts from DFAS identified these initial six different fraud schemes, based on the analysis of both case files and the data set of known fraud payments.

Subsequent validation of the choice of six different fraud types using principle component analysis and clustering along with data from a new fraud case, reduced the number of fraud types from six to four.

Therefore, the final classification, classifies fraud into Big Systematic, Small Systematic, Piggyback and Opportunistic. Details of the analysis leading to the choice of this classification scheme are given in [23].

B. CONDITIONS NEEDING IMPROVEMENT

In an effort to uncover and discourage fraud, all DoD sites responsible for trade relations with citizens, service providers, or vendors are subject to periodic visits by auditors from different offices of DFAS. At these visits a selected number of transactions are audited. The results of these audits form a growing database of current transactions from each site along with an auditor's assessment of these transactions. To choose which transactions to audit, prior to each site visit, DFAS performs an analysis of the previous eighteen months transactions. An ensemble of supervised data mining models, trained on a mix of the Knowledge Base and transactions from the specific site, and unsupervised models are used to identify potential suspect transactions. Also, duplicated transactions are identified for audit and a set of about 50 more transactions is selected randomly. At the site auditors analyze transactions, identifying specific irregularities associated with the established contracts. A detailed checklist of potential irregularities and procedural anomalies supports this in-site inspection.

The application of this checklist forms the foundation of the database of Conditions Needing Improvement (CNI's). During the process of inspection, auditors identify and record each detected irregularity associated with a particular transaction. From this detailed information, each irregularity is then classified according to its severity or CNI class. The site visit's final reports compile all the information associated with the inspected transactions and each analyzed transaction has a CNI class assigned to it. This classification of CNI's is defined as follows: Serious CNI, CNI, Irregularity, and No-CNI. In particular, the No-CNI class is assigned to those transactions that do not present any anomaly or associated irregularity and are deemed as nonfraud transactions. All the other CNI classes indicate that irregularities or procedural failings were identified, but do not constitute evidence of fraud. Further actions have to be undertaken in order to prove fraudulent practices. These audit visits constitute only a first step in the process of fraud detection.

III. MULTIPLE ADDITIVE REGRESSION TREES METHODOLOGY

A. INTRODUCTION

Multiple Additive Regression Trees (MART) is a new methodology primarily used to solve prediction problems based on the large datasets typically found in Data mining applications. Friedman [7],[8],[9] describes in detail the strategy behind this methodology which extends and improves the CART methodology and has greater accuracy than CART. It is easy to implement, automatic and maintains many of the desirable features of CART such as robustness. MART tends to be resistant to transformations of predictors and response variables, outliers, missing values, and to the inclusion of potentially large numbers of irrelevant predictor variables that have little or no effect on the response. These two last properties are of particular interest since they are two of the greatest difficulties when using transaction data to predict fraud. In this chapter a quick overview of MART is given with particular attention to interpreting the results, determining the effect of predictor variables on those results, and measuring the importance of those variables. The issues surrounding missing data are dealt with in detail in the next chapter.

B. REVIEW OF MART

MART is one of a class of methods often referred to as boosting. Boosting is a general method that attempts to "boost" the accuracy of any given learning algorithm [21] by fitting a series of models each having a poor error rate and then combining them to give an ensemble that may perform very well. In MART a series of very simple classification trees is fit, each taking very little computational effort. The MART classifier is then based on a linear combination of these trees. We describe the MART method in great detail in this section. To do so, first consider a single classification tree ([2]). For example Figure 3.1 gives the fit of a very simple tree for predicting the response variable 'FRAUD.01', (fraud = 1, nonfraud = 0), from the 2 binary predictor variables 'MILPAY' and 'INTEREST'. The terminal nodes of the classification trees (also called leaves) represent disjoint regions of the measurement space.

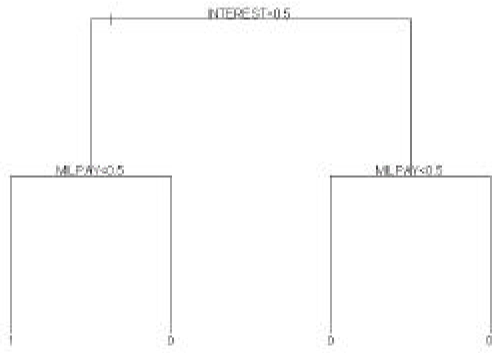


Figure 3.1 Classification tree with 4 terminal nodes

The tree in Figure 3.1 has 4 terminal nodes which splits the space of predictor variables into 4 disjoint regions {‘MILPAY’ = 0, ‘INTEREST’ = 0}, {‘MILPAY’= 1, ‘INTEREST’ = 0}, {‘MILPAY’ = 0, ‘INTEREST’ = 1}, and {‘MILPAY’ = 1, ‘INTEREST’ = 1}.

Let $x_1,..., x_L$, represent the values of L predictor variables (for a particular case or observation) and the vector $\mathbf{x} = (x_1,..., x_L)$ represent the collection of those values. The response for an observation is y. In classification, y indicates the class to which the observation belongs. For example, for the data used in Figure 3.1, L=2, and y =1 if a transaction is fraud and y =0 otherwise. The terminal nodes of a classification tree split the L dimensional space of possible predictor variables into J disjoint regions $R_j, j = 1,..., J$. The tree represents a prediction rule $f(\mathbf{x})$ for each possible value of predictor variables $\mathbf{x}$ that assigns a constant g_j for each region $R_j, j = 1,..., J$ so that

$$\mathbf{x} \in R_j \Rightarrow f(\mathbf{x}) = g_j \ , j = 1,..., J \ .$$

Thus the parameters of a classification tree which need to be estimated are the regions R_j and the corresponding predicted values for each region g_j. Let $\Theta = \{R_j, g_j\}_1^J$ represent the parameters to be estimated for the classification tree and $I(\mathbf{x} \in A)$ be an indicator function for the set A where $I(\mathbf{x} \in A)$ assumes the value 1 if $\mathbf{x} \in A$ and 0 otherwise. Then the predicted values for a tree can be expressed for predictor variables $\mathbf{x}$ as:

$$T(\mathbf{x};\Theta)=\sum_{j=1}^{J} g_j I(\mathbf{x}\in R_j) .$$

Once the R_j are estimated, the parameters g_j are estimated for each region. In a regression problem g_j is the mean of the y's whose $\mathbf{x}$'s fall in the region R_j. In classification where each observation is classified into only one of two categories, such as the fraud/nonfraud, g_j is assigned a value of 1 if the proportion of observations in R_j whose response y is 1 is greater than the proportion of observations in R_j whose response y is 0. In the example of Figure 3.1, the proportion of fraud cases in R_1 is greater than for nonfraud, so for R_1, the predicted response $g_1 = 1$ corresponds to fraud. The difficult part is determining the R_j, for which approximation algorithms exist. Hastie[11] presents a strategy based on a greedy top-down recursive partitioning algorithm to find the R_j.

In most classification problems faced by DFAS each observation is classified into $K > 2$ categories. Often, models classify transactions as nonfraud or as one of four possible types of fraud categories. Here $K = 5$. When only audited CNI transactions are used for training models, transactions are classified as Serious CNI, CNI, Irregularity or Non-CNI, giving $K = 4$. In the $K = 5$ example, MART model fit five trees. The response variable for the first tree would be $y = 1$ if the transaction is nonfraud and 0 otherwise; for the second tree the response would be $y = 1$ if the transaction is Bigsys and 0 otherwise; the third tree would have $y = 1$ if the transaction is Opportunistic and 0 otherwise; the next one $y = 1$ if the transaction is Piggyback and 0 otherwise; and for the fifth tree the response would be variable $y_5 = 1$ if the transaction is Smallsys, 0 otherwise.

In a simple boosted tree model a large number, M, trees are fit to the data. For each observation, each of the M trees "votes" on how to classify that observation. Thus for each $\mathbf{x}$ we can represent the prediction made by the boosted tree based on $f_M(\mathbf{x}) = \sum_{m=1}^{M} T(\mathbf{x};\Theta_m)$, the number of "votes" the category corresponding to $y = 1$ gets. Clearly, if the proportion of votes $\frac{f_M(\mathbf{x})}{M}$ corresponding to $y = 1$ is greater than 0.5, then $\mathbf{x}$'s class is predicted to be 1. A generic boosting algorithm can be described as follows, where $m = 1,...,M$ represents the number of trees (iterations):

```
Equally weight all the observations $(y, \mathbf{x})$ ;
For m= 1 to M
    Fit a classifier $T(\mathbf{x};\Theta_m)$ where $\Theta_m$ is the estimate of $\Theta$ at the $m^{th}$ step;
    Increase the weight of the observations which are "hard " to classify;
Define the boosted classifier as $f_M(\mathbf{x}) = \sum_{m=1}^{M} T(\mathbf{x};\Theta_m)$
```

Drucker [5] gives a good description of a boosting algorithm of this type, the AdaBoost algorithm due to Freund and Schapire. Here much like MART, the key idea is that each classifier (termed a weak learner) is trained sequentially. A subset of observations randomly selected (with replacement), from a training set, is used to train a first weak learner. As the number of trees increase, the individual training error rate increases since new weak learners have to classify more and more difficult patterns. However, the boosting algorithm shows us that the ensemble training and test error rate decrease as the number of weak learners increases. (see Figure 3.2, adapted from Drucker[5]).

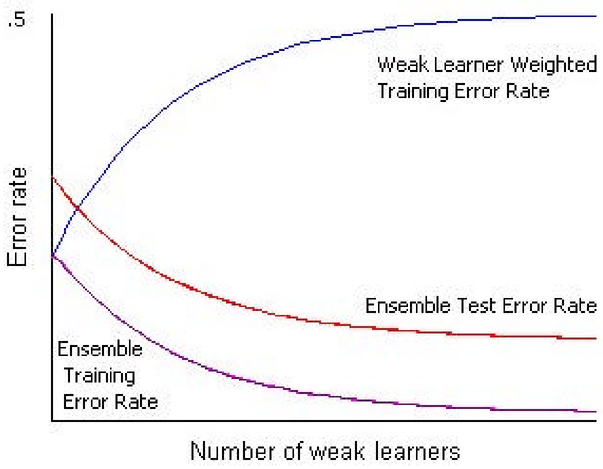


Figure 3.2 Weak learner error rate, ensemble training and test error rates

MART is a generalization of the tree boosting that attempts to increase predictive accuracy with only a moderate sacrifice of the desirable properties of trees, such as speed and interpretability. Due to the boosting process, MART produces an accurate and effective off-the-shelf procedure for data mining[11].

The MART classifier is of the form

$$f_M(\mathbf{x}) = \sum_{m=1}^{M} d_m T(\mathbf{x};\Theta_m)$$

where the additional parameters d_m, $m = 1,...,M$, are estimated sequentially at each iteration m of the MART algorithm.

Tuning parameters associated with the MART procedure are the number of iterations (individual classification trees) M and the size (number of terminal nodes or leaves) of each of the constituent trees $F_m, m = 1,...,M$.

With MART all trees are restricted to be the same size, $F_m = F$, $m = 1,...,M$. A F-terminal node classification tree is fit for each iteration. Thus, F becomes a meta-parameter of the entire boosting procedure, to be adjusted to maximize estimated performance for the data at hand.

Besides the size of the constituent trees, *F*, the other meta-parameter of the MART procedure is the number of boosting iterations (individual classification trees) *M.* Using more iterations usually reduces the training risk; however, fitting the training data too well can lead to overfitting, which degrades the risk on future predictions. A convenient way to estimate the optimal number of iterations M^* is to monitor the prediction risk as a function of M on a validation sample. The value of M that minimizes this risk is taken to be an estimate of M^*.

C. INTERPRETING MART

In general, linear combinations of trees lose the high interpretability of single decision trees. However, there are two tools that are available with MART that help interpretation of these models. This section deals with the MART contribution for helping understand boosted tree models: the measurement of relative importance of predictor variables and the use of partial dependence plots. Examples of the use of both of these tools are presented here.

1. Relative Importance of Predictor Variables

In data mining applications the input variables are seldom equally relevant. Often only a few of them have substantial influence on the response; the vast majority are irrelevant and could just as well have not been included. Ideally, a learning algorithm performance would improve when additional information is supplied by new variables. But, additional variables can interfere with other more useful ones and reduce the algorithm performance. Thus, it is often useful to learn the relative importance or contribution of each input variable in predicting the response.

Concerning the set of variables DFAS has to deal with, we are particularly interested in finding the variables that historically perform best and then use this set as a basis for future learning, recognizing that the identified set will change. Caruana *et al.* [4] states some advantages of automating the variable selection process:

- It is often difficult to determine the effects different variable combinations will have on a learning procedure. Manually selecting variables is challenging and frequently leads to inferior selection.
- It gives the system designer freedom to identify as many potentially useful variables as possible and then let the system automatically determine which ones to use.
- It allows new variables to be easily added to a system on the fly.
- In domains where the world changes, it allows the currently used variables to be those best suited to the current state of the world.
- It allows the set of variables to change dynamically as the amount of training data changes.

Friedman[9] and Hastie[11] offer a related overview about the way MART approaches variables selection, based on the concept of relative importance of predictor variables. For a single tree *T*, the measure of relevance $I_l^2(T)$ for each predictor variable x_l, $l = 1,...,L$, was proposed by Breiman *et al* [2]:

$$I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \times 1(x_l(t)) .$$

This represents the squared relative importance of variable x_l for a single J-sized classification tree T where $x_l(t)$ is the splitting variable associated with node t. The sum is defined over the $F - 1$ internal nodes of the tree T. At each such node t, one of the predictor variables x_l is used to partition the region associated with that node into two sub-regions; within each a separate constant is fit to the response values. The selected variable is the one that maximizes the estimated improvement $\hat{i}_t^2$ in squared error risk over that for a constant fit over the entire region. The sum of the squared improvements over all internal nodes for which variable x_l was chosen as the splitting variable gives its squared relative importance.

For multiple additive trees, relative importance measure for each predictor variable can be extended as being the average over all trees:

$$I_l = \sqrt{\frac{1}{M}\sum_{m=1}^{M} I_l^2(T_m)} \, .$$

Because I_l is a relative measure, in practice the value 100 is assigned to the largest value, and all the other ones are scaled accordingly.

For the case where each observation is classified into $k > 2$ categories, k boosted trees each composed of M trees is fit. For predicting the k^{th} category, $k = 1,...,K$, let T_{km}, $m = 1,...,M$, represent these M trees. Then for each set of M trees the relevance of the l^{th} predictor variable x_l in predicting the k^{th} category can be defined as:

$$I_{lk}^2 = \frac{1}{M}\sum_{m=1}^{M} I_l^2(T_{km}), \; l = 1,...,L \, , \; k = 1,...,K \, .$$

In particular the $L \times K$ matrix of these individual measures I_{lk} is very useful for identifying individual contributions. Column sums $I_{.k}^2 = \sum_{l=1}^{L} I_{lk}^2$ give the relative variable importance in predicting class k. Row sums $I_{l.}^2 = \sum_{k=1}^{K} I_{lk}^2$ represents the influence of x_l in predicting the respective classes. Averaging over all the classes, the squared overall relevance of the predictor variable x_l is given by

$$I_l^2 = \frac{1}{K}\sum_{k=1}^{K} I_{lk}^2 \, , \; l = 1,...,L \; .$$

The following description illustrates the applicability of MART to the problem of CNI prediction. A reference of commands for MART analysis inside R [8] can be found in Appendix A.

Figure 3.3 shows the results of fitting MART with 14 iterations (*M* = 14) to the OAK-Detailed data (58 predictor variables), selecting a tree-size 3 (*F* = 3) parameter. The overall misclassification rate in the OAK-Detailed set is 33.6%. Each bar represents one of the CNI classes (1 correspond to Serious CNI, 2 to CNI and 4 to No-CNI). The length of each bar indicates the fraction of the test set observations that were misclassified within each class.

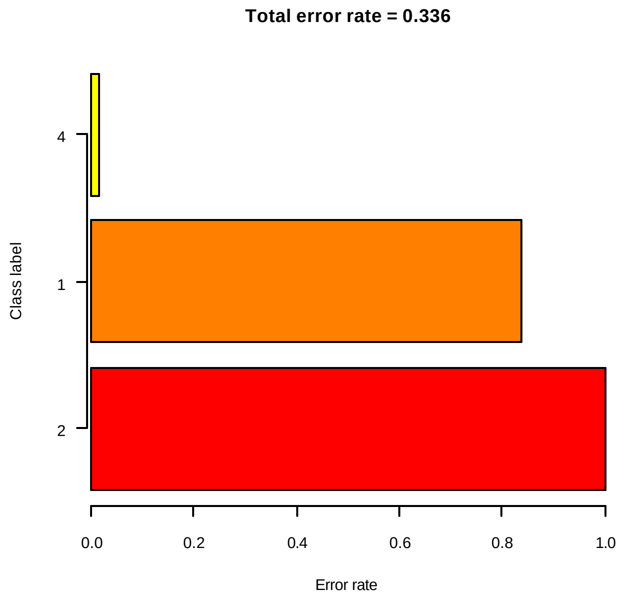


Figure 3.3 Total error rate for CNI classification of OAK-Detailed data with MART(*M*=14,*F*=3).

The horizontal barplot shows that class 2 (CNI) is the most difficult to classify, presenting 100% of the cases misclassified, while class 4 (No-CNI) presents a relatively low misclassification rate. Class 1 (Serious CNI) also presents a considerable misclassification rate (80%). Figure 3.4 gives the detailed misclassification error rate for each CNI class. The CNI class 1 (Serious CNI) and class 2 (CNI) show misclassifications with class 4 (No-CNI). In particular, class 2 (CNI), which shows a misclassification of 100%, is being misclassified with CNI class 1 (10%) and CNI class 4 (90%).

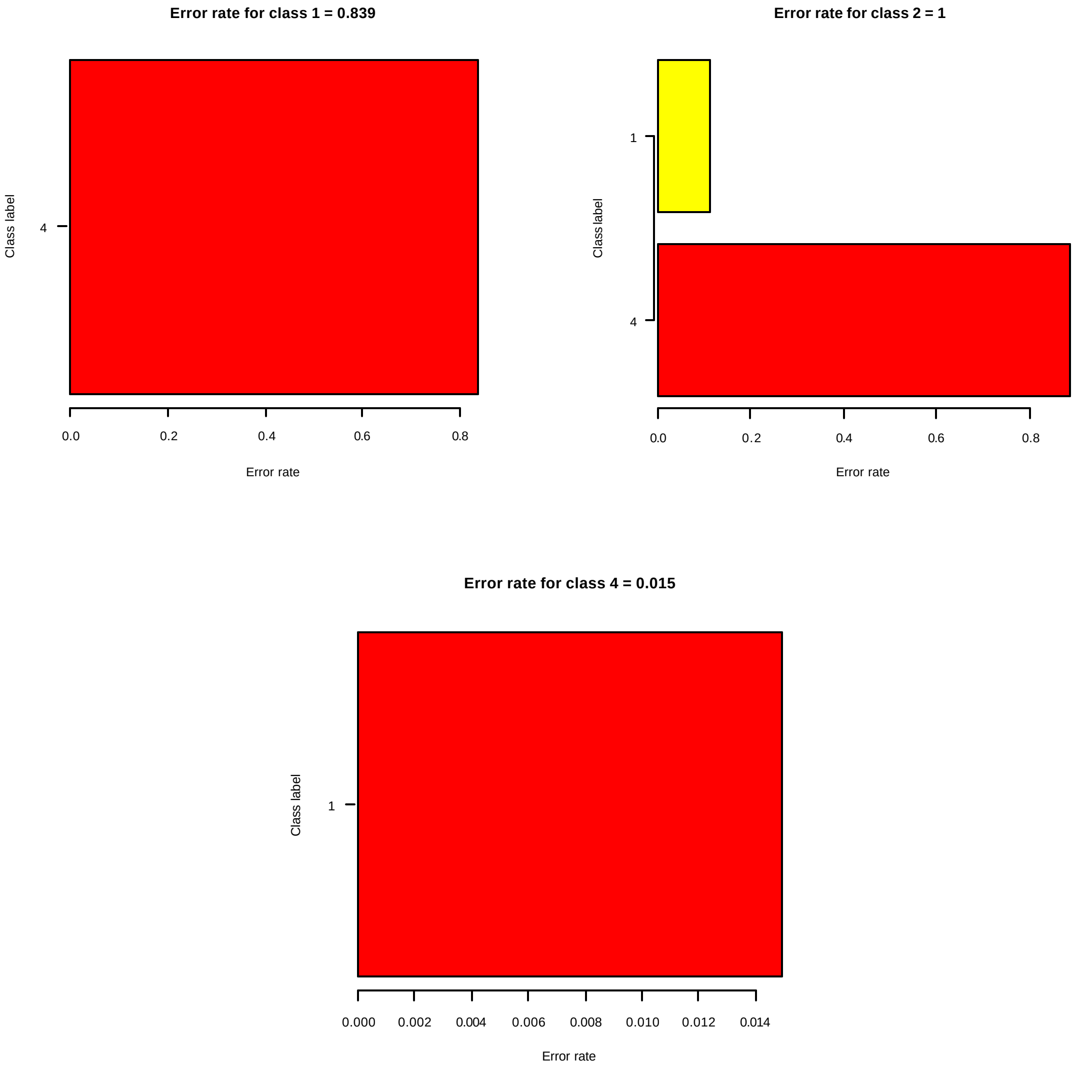


Figure 3.4 Misclassification error rate for CNI classes; in classification of OAK-Detailed data with MART(*M*=14, *F*=3).

For each CNI category MART constructs fourteen trees of 3 leaves, and therefore 2 splitting nodes, identifying the twenty-four (24) important variables on differentiating, with lowest error rate, all CNI classes (Figure 3.5).

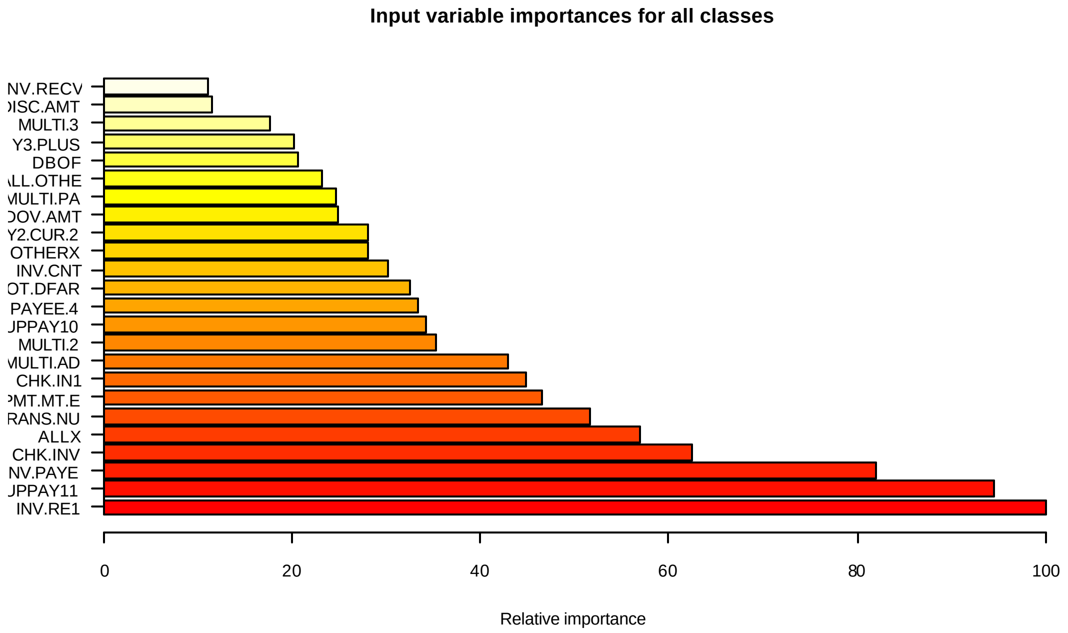


Figure 3.5 Relative importance of predictors for CNI classification of OAK-Detailed data with MART(*M*=14, *F*=3).

As an example of how MART measures a variable's relative importance for each classification category: consider trying to predict a response variable with four categories (ex. Fraud), using trees of size two (*F*=2; one splitting node – such a tree is called a "stump") and one iteration (*M*=1). It follows that the final ensemble will have one stump for each category. In this way MART identifies the most significant variables responsible for separating each category's observations from the other classes. With this illustrative example, we would end up with at most four variables identified as important, one for each generated stump; this particular situation presents the predictor's squared measure of relative importance, given directly by its maximum estimated improvement $\hat{i}_t^2$ in squared error risk.

Using the same data and parameters from the initial example, the predictor variable importance for each CNI class is presented in Figure 3.6.

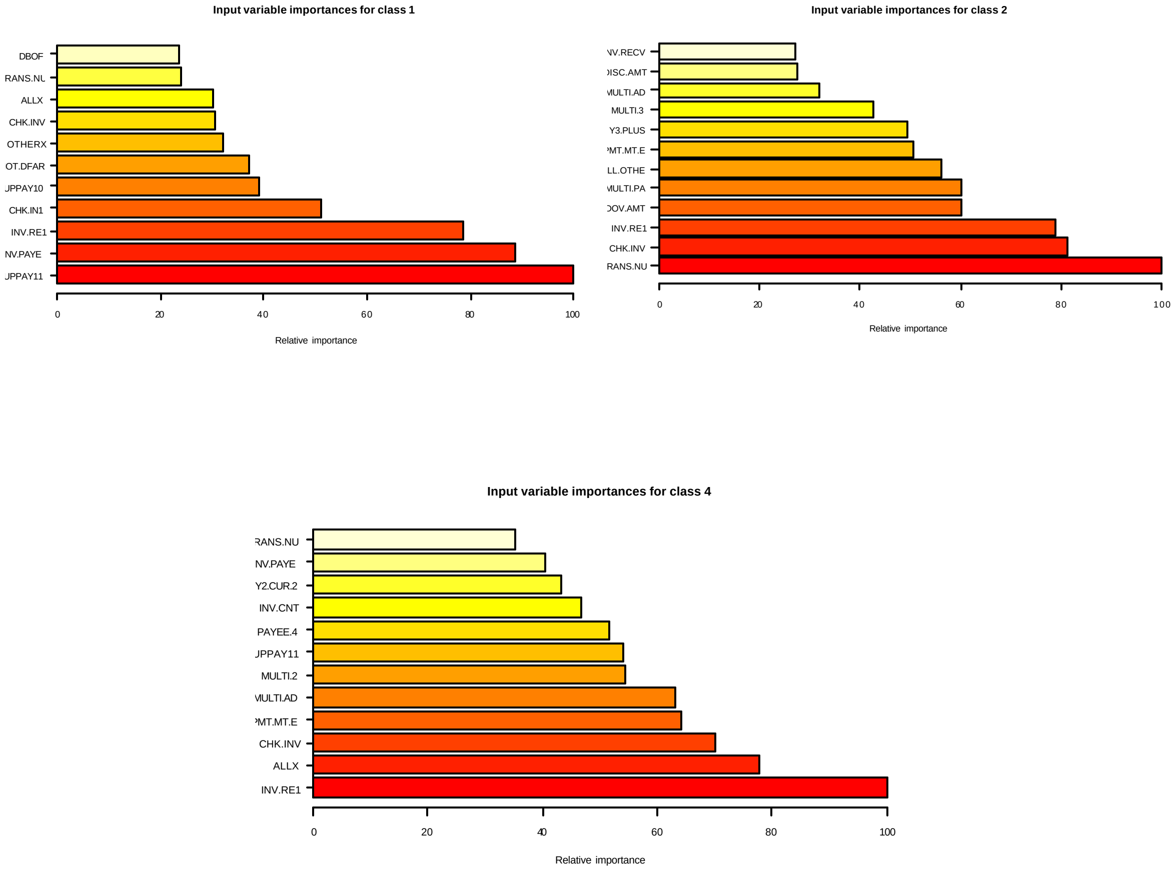


Figure 3.6 **Predictor variable importance for each of the three CNI class for OAK-Detailed data with MART(*M*=14, *F*=3).**

Besides a variable's relative importance, it might be interesting to know, for a given predictor variable, which classes it helps identify best. Following the example of classification for OAK-Detailed data with MART(M=14, J=3), one can see in Figure 3.7 the contribution of the predictor variable 'ALLX' in classifying class 2 and class 4 CNI. The contribution of a predictor variable is presented in a rescaled percentage bar plot. The predictor variable 'ALLX' contributes more towards separating class 4 than in separating class 2, and no contribution is shown in separating other class.

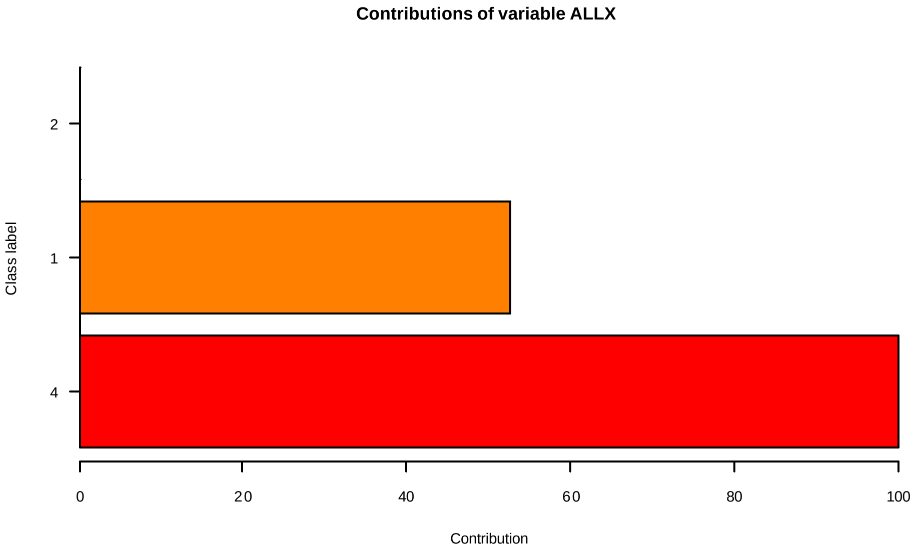


Figure 3.7 Contribution of predictor variable 'ALLX' for CNI's classification for OAK-Detailed data with MART(*M*=14, *F*=3).

2. Partial Dependence Plots

After identifying the most relevant variables, one might be interested in attempting to understand the nature of the dependence of the boosted tree model on their joint values.

The analysis of partial dependence plots of the boosted tree approximation on selected variables subsets can help to provide a qualitative description of its properties. In particular it is expected that those subsets whose effect on the model is approximately additive or multiplicative will be most helpful.

With a *K*-class classification problem, *K* separate MART models are fit, one for each class. The response variable y assumes values in the unordered set $G = \{G_1, ..., G_k\}$, and a particular classifier $G(\mathbf{x})$ taking values in G can be defined knowing the class conditional probabilities $p_k(\mathbf{x}) = \Pr(y = G_k \mid \mathbf{x})$, $k = 1, ..., K$ as being

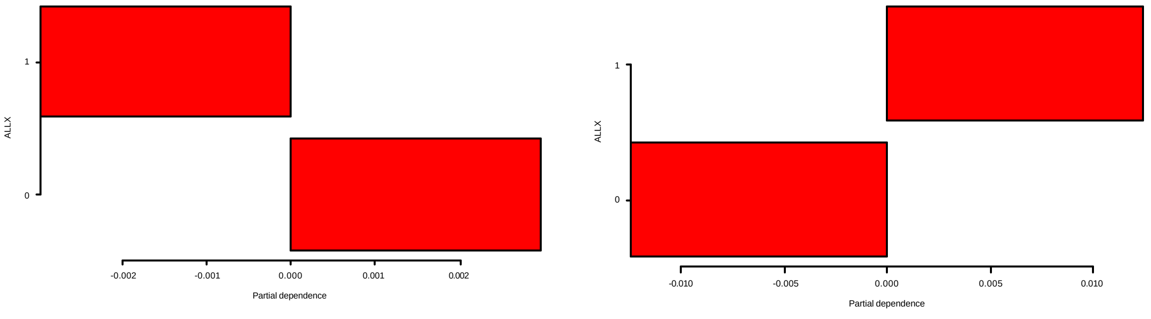
$$G(\mathbf{x}) = G_k \text{ where } k = \underset{1 \le l \le K}{\operatorname{argmax}}\, p_l(\mathbf{x}).$$

where one only needs to know the largest $p_k(\mathbf{x})$. In particular, for data mining problems we might be more interested in the class probabilities $p_k(\mathbf{x})$, $k = 1,\ldots,K$ themselves, rather than in performing a class assignment. Let $f_{kM}(\mathbf{x})$ be the MART classifier for class k, then $p_k(\mathbf{x})$ is usually taken to be:

$$p_k(\mathbf{x}) = \frac{e^{f_{kM}(\mathbf{x})}}{\sum_{l=1}^{K} e^{f_{lM}(\mathbf{x})}}, \ k = 1,\ldots,K \ .$$

which ensures that $0 \le p_k(\mathbf{x}) \le 1$ and that they sum to one. Larger values of $f_{kM}(\mathbf{x})$ imply a higher probability of observing class *k* associated with the predictor variable x. Partial dependence plots of each boosted tree's model on a particular subset of predictor variables most relevant to class *k* provide information on how the input variables tend to increase or decrease the odds of observing that class.

The partial dependence of the model in each predictor variable can be visualized, providing insight into what values of those variables tend to increase or decrease the odds of observing that class. If variables are categorical variables, the partial dependence plot is a horizontal bar plot. The bars are ordered, bottom to top, in ascending categorical value and the length, positive and negative, of each corresponding bar represents the value of the partial dependence function for that variable value. Plots are centered to have mean partial dependence of zero. Following the example of classification for OAK-Detailed data with MART(M=14, J=3), Figure 3.8 presents the partial dependence of class 4 (No-CNI) and class 1 (Serious-CNI) on the values of the predictor variable 'ALLX'. The illustration in Figure 3.8 is conforming to Figure 3.7; variable 'ALLX' increases the odds of observing class 4, Figure 3.8 (b), more than class 1, Figure 3.8 (a).



(a) Serious-CNI

(b) No-CNI

Figure 3.8 Partial dependence plots of 'ALLX' for class 1 and class 4 for OAK-Detailed data with MART(*M*=14, *F*=3).

Figure 3.9 shows the partial dependence of class 1 (Serous-CNI) on joint values of the predictor variables 'DBOF' and 'ALLX', i.e., the partial dependence of the model for class 1 on 'ALLX', conditioned on the variable 'DBOF'.

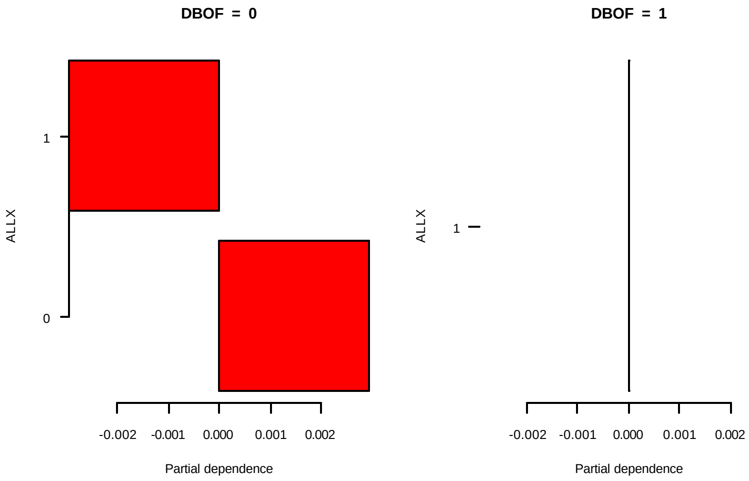


Figure 3.9 Partial dependence of class 1 on joint values of 'ALLX' and 'DBOF' for OAK-Detailed data with MART(*M*=14, *F*=3).

The nature of the partial dependence plots of the model on the joint values of two predictor variables depends on the type of the variables. If both are real-valued, then a perspective mesh plot representing the value of the partial dependence function for joint values of the variables is produced. If both variables are categorical, a series of bar plots is produced. Each bar plot represents the partial dependence of the model on the plotted variable, conditioned on the corresponding value of the other variable.

IV. KNOWLEDGE BASE ANALYSIS

A. INTRODUCTION

The Knowledge Base constitutes the repository of known fraud transactions. It contains information about each identified fraud transaction compiled in a common format. This common format has fields containing original information and also new fields, indicators developed by subject matter experts from DFAS, DMDC and the DODIG. When matched against transaction information these indicators help to identify anomalies in the data, expose fraud or locate internal control weaknesses.

A major concern is the presence of a significant number of missing values in the Knowledge Base. This chapter focuses on the missing value problem and discusses MART's approach to this problem. It also addresses the issue about whether the classification models, based on the historical Knowledge Base, are really capturing fraud and nonfraud patterns in transactions or whether they are classifying some other feature that differentiates transactions.

B. ANALYSIS OF KNOWLEDGE BASE

1. Data

The Knowledge Base analysis is performed using the following set of variables, explained in Appendix B:

Table 4.1 Variables Set for Knowledge Base Analysis

DOV.AMT	DISC.AMT	TRANS.NU	INV.AWAR	INV.RECV	CHK.AWAR	INV.RE1	CHK.INV
CHK.IN1	MILPAY	DBOF	OTHERX	UNUSUAL	ALLX	Y1.PRIOR	Y1.CUR
Y2.CUR.1	Y2.CUR.2	Y3.PLUS	ALL.OTHE	ENHANCE	STE	POBOX	INV.PAYE
INV.CNT	DOVAMT.2	DOVAMT.1	AVG.5K	PAYEE.4	MULTI.PA	MULTI.AD	INV.SEQ
PMT.FREQ	PMT.FR1	TINS	MULTI.TI	MULTI.2	MULTI.3	MULTI.4	FEW.PYMT
MISC.OBL	DUPPAY10	DUPPAY11	NOT.DFAR	NUMADR.K	NUMADREE	NUM.EE.K	MANIND.A
MANIND.M	PMT.MT.E	PMT.MT.D	PMT.TP.C	PMT.TP.F	PMT.TP.P	PMT.PR. F	PMT.PR.P
PMT.PR.R	PPA.XM.C	PPA.XM.E					

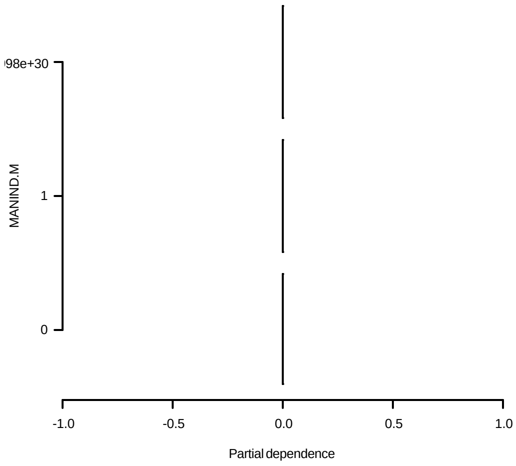
2. Missing Values Analysis

The problem of missing values in analysis is an old one [18]. The waste of data and the potential for misleading results which can result from casewise deletion of missing values has prompted the development of different techniques for dealing with missing values in classification [15]. A summary of methods for dealing with missing values is presented in Fujikava [6]. Often missing values are filled in (imputed). Properly imputing missing values in data can help in reducing the error rate of the learned concepts. However, imputation must be done with care; otherwise it will introduce excessive noise [19]. After the missing values are replaced by imputed values, the new values are typically treated as if they were actually observed. The relative ambiguity associated with the imputation process itself introduces additional uncertainty into estimates and response predictions. As suggested in Hastie *et all* [11], one can measure this additional uncertainty by doing multiple imputations and hence creating many different training sets. A predictive model for the response variable can be fit to each training set so that the variation across training sets can be assessed.

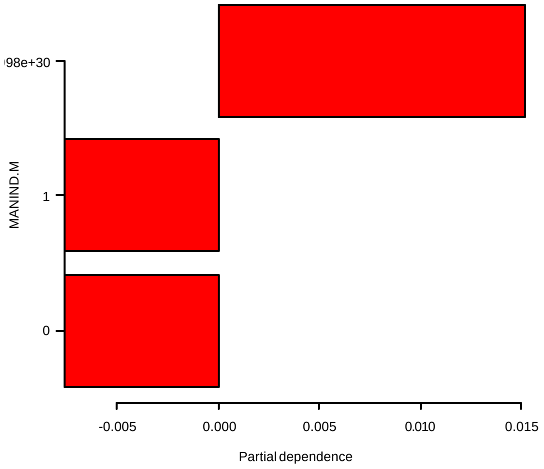
Hastie *et all* [11] recommends two particular approaches in order to deal with the missing values problem for tree-based models. The first more general approach and the approach used by CART is the construction of surrogate variables. When considering a predictor for a split, only the observations for which that predictor was not missing, were used. Having chosen the best (primary) predictor and split point, a list of surrogate predictors and split points is produced. The first surrogate is the predictor and corresponding split point that best mimics the split of the training data achieved by the primary split. The second surrogate is the predictor and corresponding split point that does second best, and so on. When sending observations down the tree either in the training phase or during prediction, the surrogate splits are used in order if the primary splitting predictor is missing. Surrogate splits exploit correlations between predictors to try to alleviate the effect of missing data. The higher the correlation between the missing predictor variable and the other predictor variables, the smaller the loss of information due to missing values. In particular, CART is an example of one learning algorithm that deals effectively with missing values through surrogate splits.

The second approach suggested by Hastie *et all* [11] is applicable to categorical predictors and defines missing values as a new category or extra class. Exploring this particular technique it might be discovered that observations with missing values for some measurement behave differently than those with nonmissing values. This approach relies on the learning algorithm to deal with the missing values in its training phase. This approach is adopted by MART methodology.

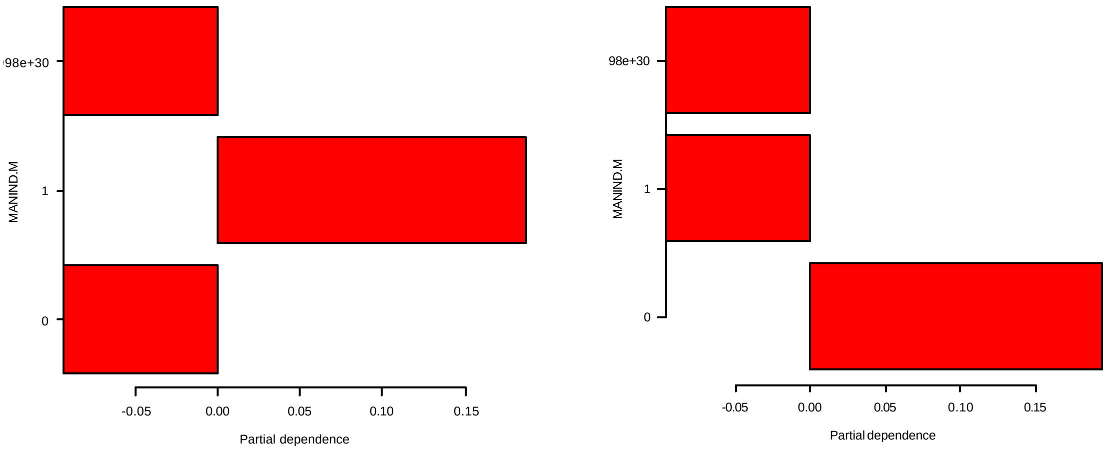
Missing values in MART are handled as an extra class inside each effected predictor variable. A considerably high value (9E+30 as default; the value 98+30 present in the plot is the result of a display truncation of 8.99999998E+30) is assigned to each missing value; in this way, the process of classification will consider missing values as a part of the classification model. The missing value classes will contribute to maintaining those observations as part of the classification process, influencing the accuracy of the classification model in the prediction process.



(a) Bigsys



(b) Opportunistic



(c) Piggyback *(d) Smallsys*

Figure 4.1 Partial dependence plots of 'MANIND.M' for fraud classes in KB.

Figure 4.1 gives an example of singleplots (generated with R 1.5.0) from the Knowledge Base analysis that illustrate the way MART deals with missing values present in the predictor variable 'MANID.M'. In particular, Figure 4.1 *(b)* gives some insights about the way missing values contributes to increasing the odds of observing fraud class 2 (Opportunistic). A contrasting effect can be seen in Figure 4.1 *(c)* and *(d)* where missing values are as important as MANIND.M category 0 for observing fraud class 3 (Piggyback), and as 'MANIND.M' category 1 for observing fraud class 4 (Smallsys). In plot (c), category 1 of 'MANIND.M' assumes capital importance in increasing the odds of observing fraud class 3, while in plot *(d)* it is 'MANIND.M' category 0 that most contributes to increasing the odds of observing fraud class 4 (Smallsys). Plot *(a)* shows that no partial dependence exists between fraud class 1 (Bigsys) and the predictor variable 'MANIND.M'.

a. Missing Data Patterns

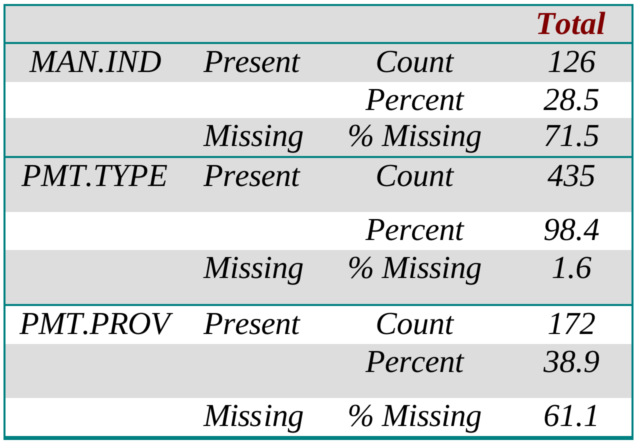
The first issue in dealing with the missing value problem is determining whether the missing data mechanism has distorted the observed data, since the knowledge of the mechanisms that led to certain values being missed constitutes a key element in choosing an appropriate analysis and interpreting the results. Little and Rubin [14] divide these mechanisms into three categories: Missing Completely at Random

(MCAR), Missing at Random (MAR), and Nonignorable. Missing data is considered to be MCAR when the mechanism that produces the missing data is not related to the value that should have been observed for that data point. Generally, one can test whether MCAR conditions can be met by comparing the distribution of the observed data between the subset with missing values and the subset without missing values. Missing data is considered to be MAR when the missingness depends on some of the variables in the analysis, but conditional on those variables, is not related to the value that should have been observed for that data point, i.e, the mechanism resulting in its omission is independent of its (unobserved) value [11]. MCAR is a stronger assumption than MAR; most imputation methods rely on MCAR for their validity. Lastly, missing data is considered to be nonignorable when missingness is nonrandom and is not predictable from any one variable in the database. Typically, this type of missing data is the hardest condition to deal with, but unfortunately, the most likely to occur as well.

b. Nonrandom Pattern of Missing Values

DFAS recognizes that one of the greatest problems in the Knowledge Base is the high rate of missing values for particular predictor variables. Table 4.2 shows the most relevant information about missing value variables with a missing percentage greater than 1%.

Table 4.2 Missing values in Knowledge Base predictor variables.



			Total
MAN.IND	*Present*	*Count*	*126*
		Percent	*28.5*
	Missing	*% Missing*	*71.5*
PMT.TYPE	*Present*	*Count*	*435*
		Percent	*98.4*
	Missing	*% Missing*	*1.6*
PMT.PROV	*Present*	*Count*	*172*
		Percent	*38.9*
	Missing	*% Missing*	*61.1*

A two-sample *t* test is one way to check if data are missing completely at random. If the values of a variable are MCAR, then other quantitative variables should have roughly the same distribution for cases separated into two groups based on pattern: missing or present.

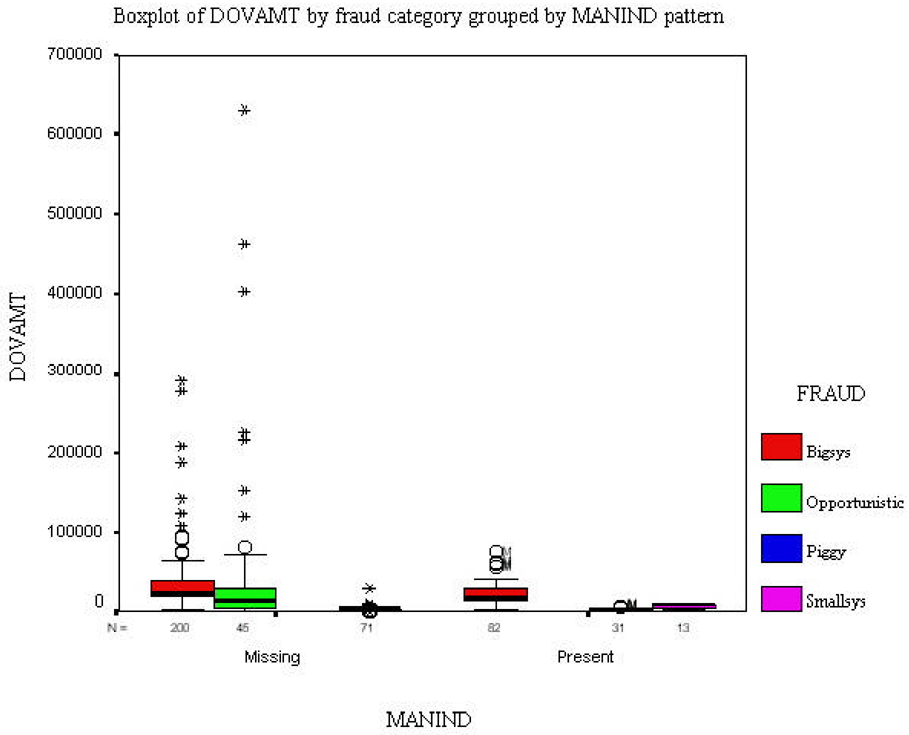


Figure 4.2 Boxplot of DOVAMT by fraud category grouped by 'MANIND' pattern.

Figure 4.2 shows the Disbursing Office Voucher Amount – 'DOVAMT' by fraud category grouped by Manual Indictor 'MANIND.' Within each fraud class, means of 'DOVAMT' are slightly higher for missing 'MANIND' values. Despite the reduced number of cases in missing 'PMTTYPE' pattern, the mean value of 'DOVAMT' is significantly higher for observations missing the 'PMTTYPE' pattern than for those in which it is present, as shown in Figure 4.3.

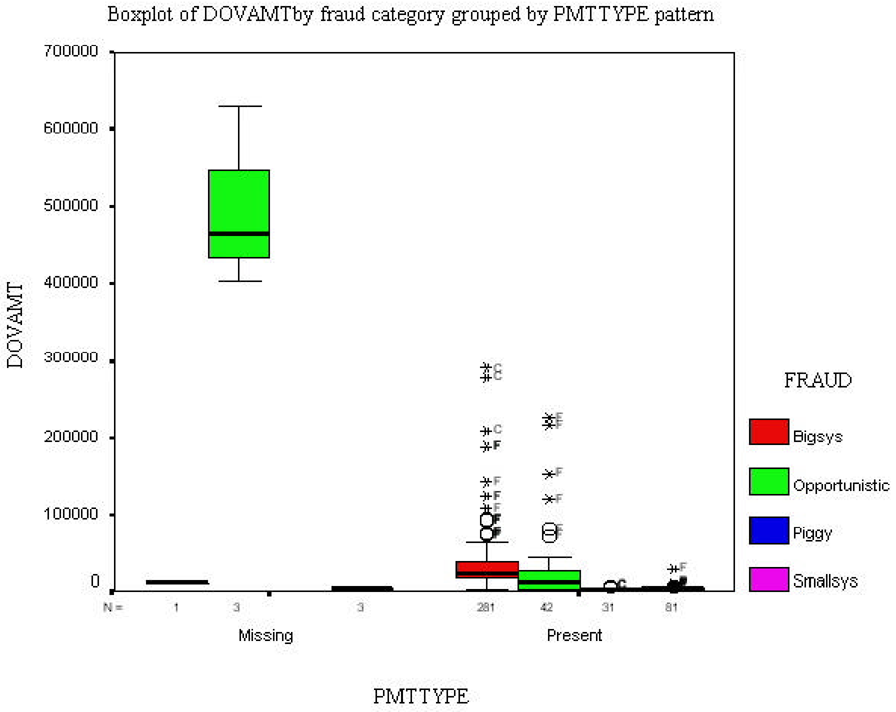


Figure 4.3 Boxplot of 'DOVAMT' by fraud category grouped by 'PMTTYPE' pattern.

A missing indicator variable that indicates whether the value of the variable is present or missing is created for each variable. Welch's modification of the Student's *t* Test for differences in means with unequal variances is used with groups formed by indicator variables. The groups are determined by whether the indicator variable is coded present or missing. The *t* statistic, counts of missing and nonmissing values, number of degrees of freedom, and means of the two groups are displayed in Appendix D.

Results of the *t* test confirm that average 'DOVAMT' is significantly higher when 'MANIND' is missing than when it is present. For each quantitative variable, the unequal variances *t* test is performed, comparing the means for those variables (row) with at least 1% missing data; when only a few values are missing, the *t* statistics are not informative.

Both the size and location of bolded *t* values in Table 4.3 confirm earlier observations that 'MANIND' and 'PMT.PROV' have different nonrandom patterns of missing data.

Table 4.3 Unequal Variance *t* tests – *t* values

	DOV.AMT	*DISC.AMT*	*TRANS.NU*	*INV.AWAR*	*INV.RECV*	*CHK.AWAR*	*INV.RE1*	*CHK.INV.*	*CHK.IN1*
MAN.IND	***-4.4***	***3.3***	*-.9*	***10.9***	***10.1***	***9.1***	***-3.6***	***-6.9***	***-6.4***
PMT.TYPE	*-1.8*	***3.2***	***2.3***	***2.9***	***3.1***	***3.5***	***2.3***	***5.5***	***3.8***
PMT.PROV	*-1.8*	***3.0***	***2.4***	***6.3***	***6.2***	***5.4***	*-1.0*	***-3.1***	***-4.0***

Figure 4.4 shows a profile of the mean differences of those variables when 'MANIND' is missing and present. All the variables are transformed to z scores in order to assure comparability. When 'MANIND' is missing, the means of all quantitative variables are lower than when it is present.

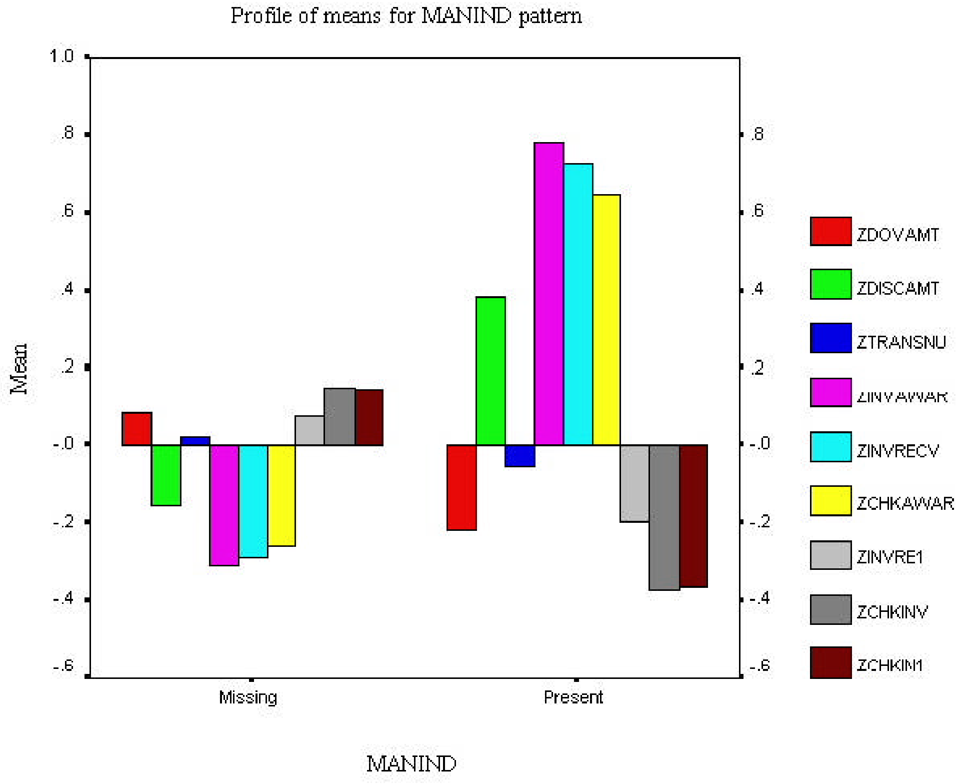


Figure 4.4 Profile of means for 'MANIND' patterns

In conclusion, it can be said that missing values in the Knowledge Base are not missing completely at random but rather present a nonrandom pattern.

3. Analysis With MART

The following analysis offers an overview over the Knowledge Base data-structure using MART. This initial analysis, based on the MART (M=179, J=2) model, highlights some issues such as fraud classification structure inside the Knowledge Base and the way MART deals with missing values.

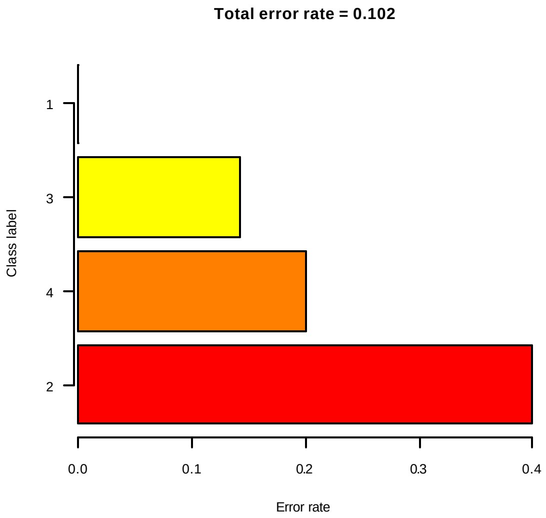
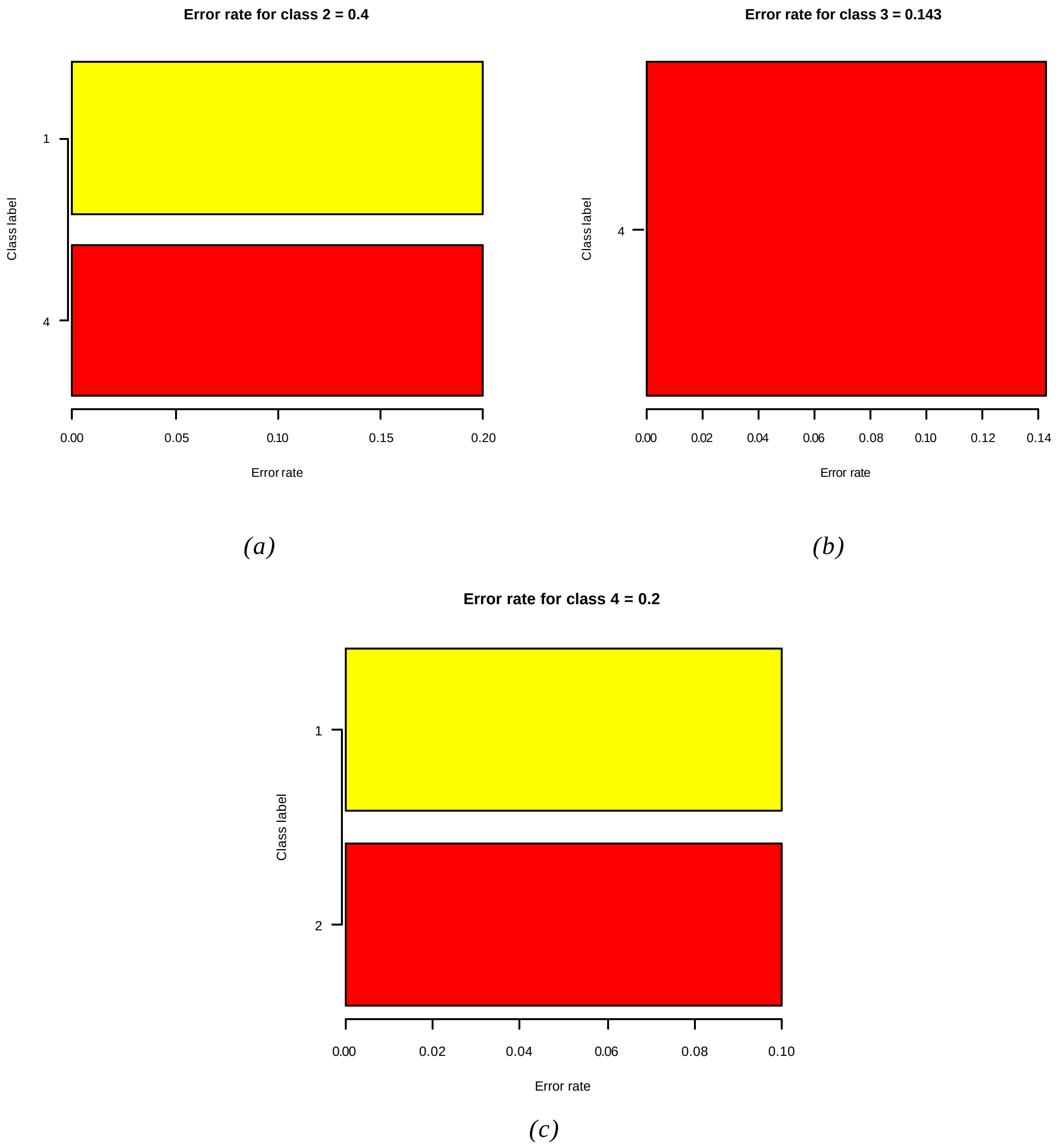


Figure 4.5 Total misclassification error

The misclassification risk structure in Figure 4.5 shows that total error rate (10.2%) is associated with fraud class 2 (Opportunistic), 3 (Piggy) and 4 (Smallsys) misclassifications.

Exploring the misclassification associated with each fraud class, Figure 4.6 *(a)*, *(b)* and *(c)* describes the error rate associated respectively with fraud class 2 (Opportunistic), 3 (Piggy) and 4 (Smallsys).

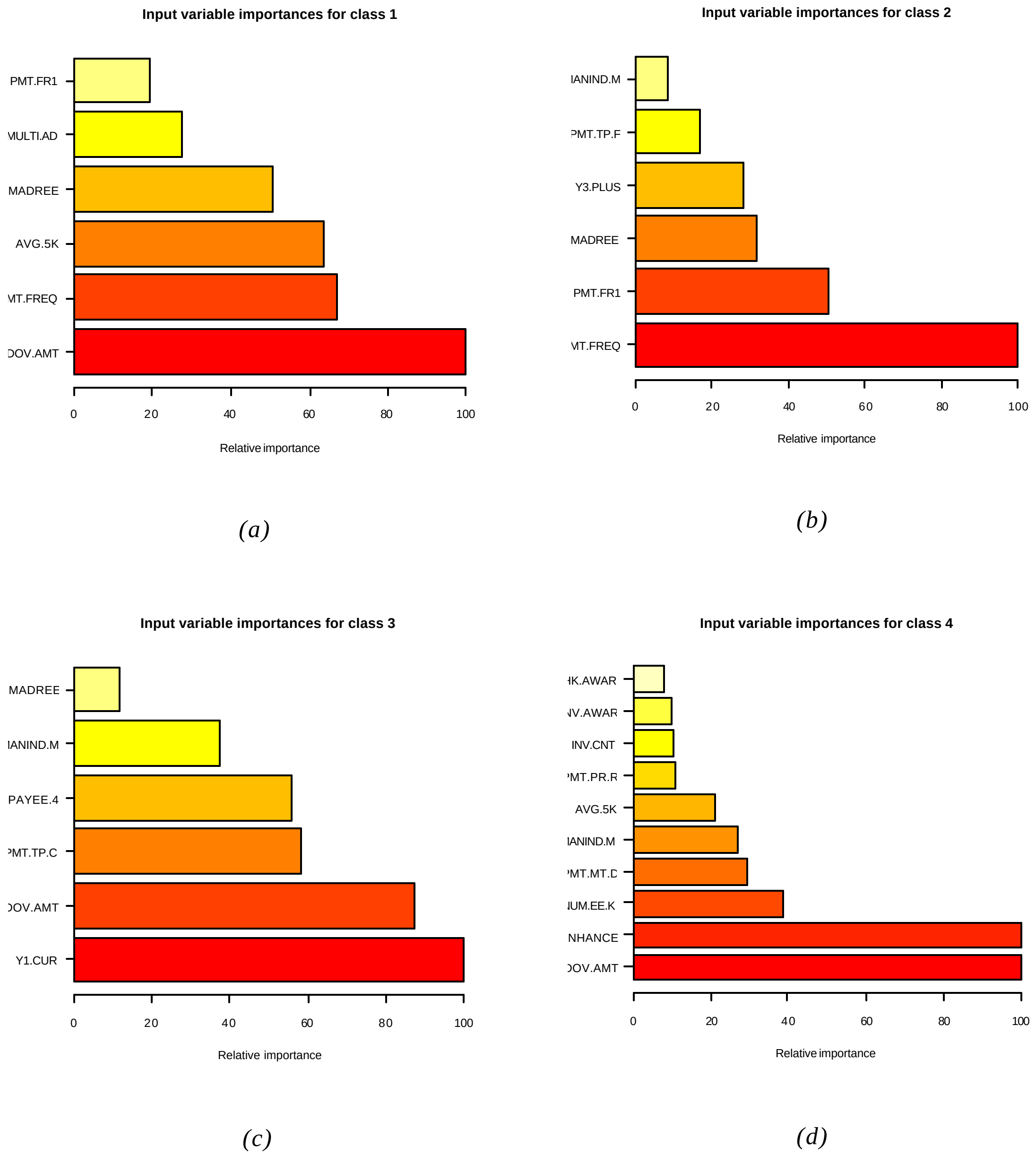


(a) *(b)*

(c)

Figure 4.6 Fraud classes misclassification error.

The next set of plots in Figure 4.7 shows the relative important variables in classifying each fraud class inside the Knowledge Base.



(a)

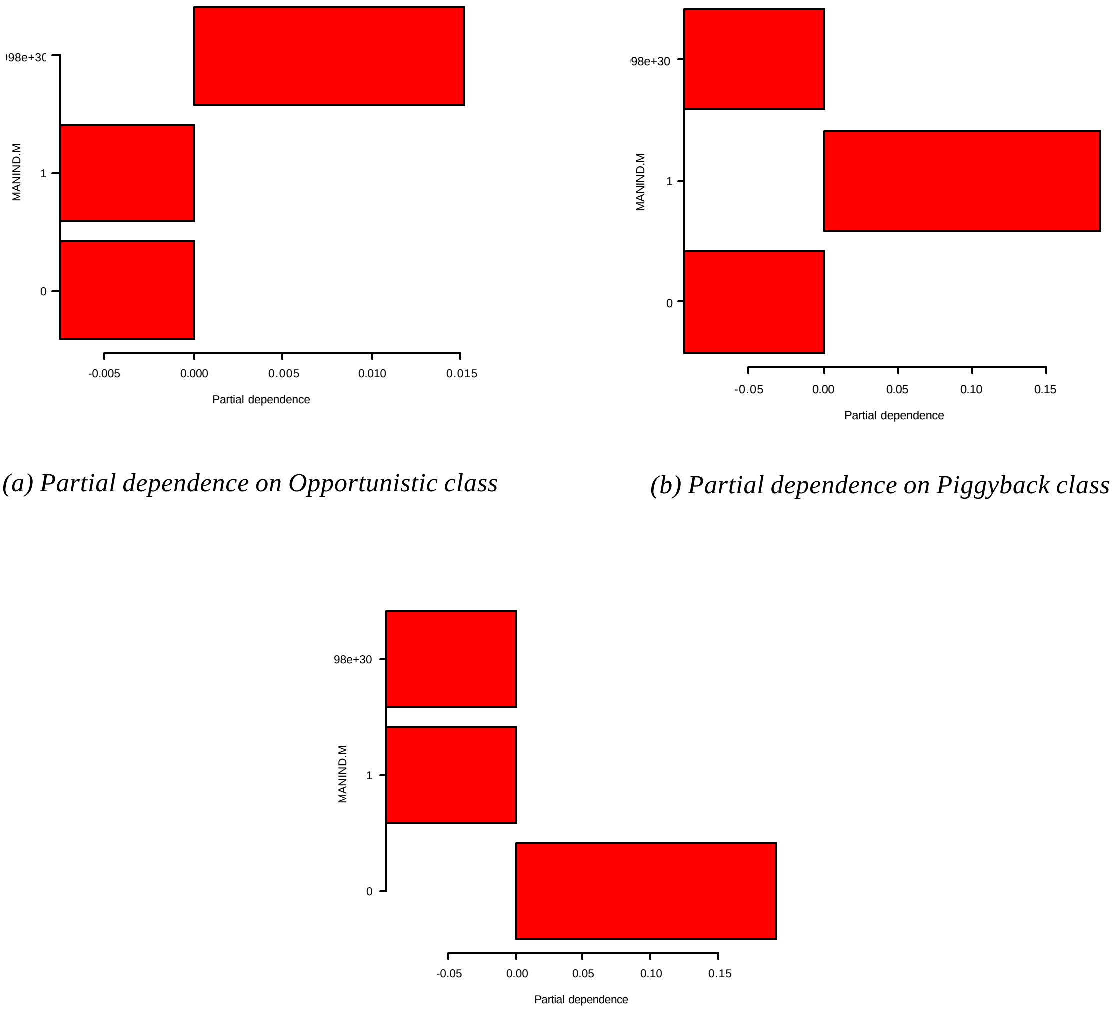
(b)

(c)

(d)

Figure 4.7 Important variables for classifying different fraud categories.

Figure 4.8 explores partial dependences of the missing valued variable, ‘MANIND’. Variable ‘MANIND.M’ has partial dependences on fraud class 2 (Opportunistic), class 3 (Piggy) and class 4 (Smallsys). This is not surprising since this variable is present in all sets of important variables except for fraud class 1 (Bigsys). For class 2 in particular (Figure 4.8 *(a)*), the partial dependence is mostly due to the missing values class. Notice that ‘MANIND.A’ and ‘MANIND.M’ present the same missing value structure.



(a) Partial dependence on Opportunistic class

(b) Partial dependence on Piggyback class

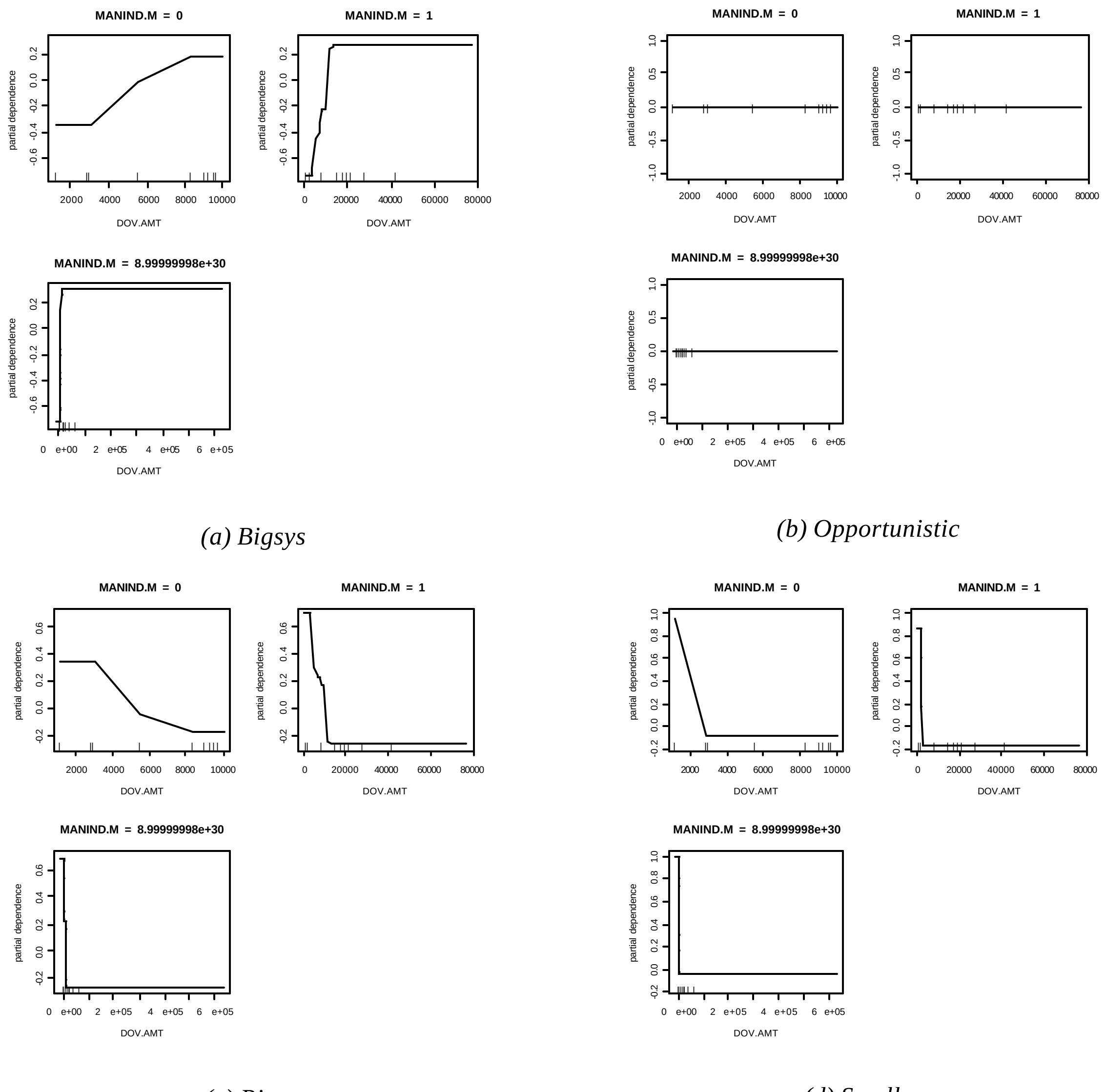
(c) Partial dependence on Smallsys class

Figure 4.8 Partial dependence plots of ‘MANIND.M’ on fraud classes

Further conclusions can be drawn from similar partial dependence plots, in particular for partial dependence of ‘PMT.TP.F’ on fraud class 2, ‘PMT.TP.C’ on class 3 and ‘PMT.MT.D’ and ‘PMT.PR.R’ on class 4. The most significant issues are missing value related and it can be noticed in all these variables the presence of related missing values dependence on classifying each related fraud class.

Special attention ought to be paid to the conditional dependence of predictors in addition to the singular partial dependence of fraud classification on identified important predictors.

Figure 4.9 shows conditional dependence of fraud classes on predictors 'DOVAMT' and 'MANIND.M'. Plots show the partial dependence of each fraud class on variable 'DOVAMT', conditioned on each value 'MANIND.M' assumes, including the missing value category. As expected in plot Figure 4.9 *(b)* one can see no dependence of fraud class 2 on 'DOVAMT' since was not identified as important for classifying Opportunistic fraud.



(a) Bigsys

(b) Opportunistic

(c) Piggy

(d) Smallsys

Figure 4.9 Paired dependence plots of 'DOVAMT' conditioned by 'MANIND.M' on fraud.

Plot Figure 4.9 *(a)*, *(c)* and *(d)* show respectively the partial dependence of fraud class 1 (Bigsys), class2 (Opportunistic) and class 4 (Smallsys) on 'DOVAMT' when conditioned by each 'MANIND.M' category, including the missing value category. It can be seen that the partial dependence on 'DOVAMT' conditioned by missing values and the partial dependence on 'DOVAMT' conditioned by 'MANIND.M' follow a paired similar pattern for class 3 (Piggy) and class 4 (Smallsys) fraud classes.

The hash marks at the base of each plot identify the deciles of the data distribution of the corresponding variables.

By repeatedly applying the procedures *singleplot* and *pairplot* to generate single partial dependences and conditional dependences plots, one can graphically examine the predictive relationship between predictor variables and fraud.

4. Imputing Missing Values

Because imputing missing values is such a common technique, it is important to compare the results of imputations with the results of MART's method of handling missing values. The technique used in this section to impute missing values is based on a voting system in which missing values in some variable are estimated by the remaining set of predictor variables. MART is the tool used for this prediction.

A separate MART model is fit for each variable with identified missing values. In each model that variable is taken to be the response. All the other variables constitute the set of predictors used to train the classification model.

For each missing-valued variable in the Knowledge Base only those observations without missing values are used to train the MART model; the remaining set of observations constitute the classification target set (those for which missing values are imputed or estimated with the MART model).

Two analyses are performed, one without the fraud variable included in the set of the predictor variables, and a second one including this particular field. The basis of this approach resides in the elementary principle of imputing a missing value based on the contribution of other variables. This approach directs the attention of strong predictor candidates to those variables affected by missing values, here called weak predictor candidates. The concept of weak and strong predictor candidates resides in the degree of present or absent missing values. The objective is to strengthen the set of predictor variables, reducing the degree of fuzziness in predicting fraud.

For each variable with missing values, a “best” model, the model with smallest error rate and tree size is selected. Figure 4.10 summarizes the ensemble dimension and tree size for the best model for each variable with missing values.

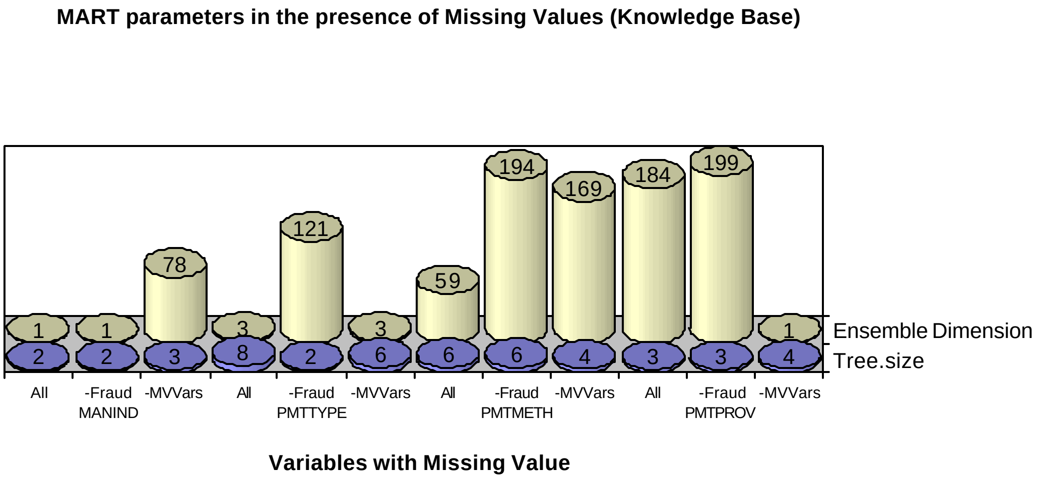


Figure 4.10 Selecting the best model for missing values analysis

We note that because the variables with missing values are categorical, imputation with MART yields predicted probabilities that the missing value falls into a particular class. For example the variable ‘PMTTYPE’ takes three possible values ‘PMTTYPE’=C, ‘PMTTYPE’=F and ‘PMTTYPE’=P. Predicted probabilities for the three possible categories for PMTTYPE and for seven transactions where PMTTYPE is missing are given in Table 4.4. This particular example is useful to illustrate how different prediction probabilities are used to select the final classification. The missing value is entered as the category with the largest predicted probability. The different discriminative power of prediction probabilities associated with cases 1 and 7 are illustrative of the degree of fuzziness one prediction can have.

Table 4.4 Predicted ‘PMTTYPE’ categories using swarm technique in the Knowledge Base.

Observation	PMTTYPE=C	PMTTYPE=F	PMTTYPE=P
1	**0.1015390**	**0.6949781**	**0.2034829**
2	0.1554473	0.5076495	0.3369032
3	0.1013682	0.6862710	0.2123609
4	0.1253940	0.6723630	0.2022430
5	0.1253940	0.6723630	0.2022430
6	0.1253940	0.6723630	0.2022430
7	**0.2237105**	**0.3962993**	**0.3799900**

Finally, Table 4.5 summarizes the results of this analysis. Three different methods for handling missing values are compared based on their misclassification risk. The first method applies MART relying on MART’s default methodology for treating missing values. In the second method, missing values are guessed using common sense and no systematic methodology. The third method imputtes missing values using MART.

Table 4.5 Comparison of missing values treatments for Knowledge Base.

Knowledge Base	Best Misclassification risk	MART(M, F)
MART	0.0114	(362, 3)
Guessing	0.0114	(271, 2)
Imputting Missing Values	0.0114	(498,2)

The result of this study shows that no advantage is achieved when we proceed with missing values imputation. The effort of trying to fix the nonrandom missing value patterns in the Knowledge Base with various imputation methods does not result in models which perform better than MART.

C. THE EFFECT OF MISSING VALUES IN KB ON PREDICTING FRAUD

Here we inspect the degree to which the Knowledge Base cases training models to predict fraud. The main question resides in the fact that the mix of historical fraud cases from the Knowledge Base with current transactions might not be classifying transactions as fraud and nonfraud, but classifying transactions according to another feature that differentiates the historical fraud transactions from nonfraud transactions, such as differences in business practices in the two groups. Variables with missing values in the Knowledge Base are indicative of such changes. We cannot answer this question directly. To shed light on this issue we ask the related question: whether the differences in the proportion of fraud cases classified using the Knowledge Base at the six different sites, Dayton, Oakland, San Antonio (CAPS), San Antonio (IAPS), San Diego and Pensacola are different.

The null hypothesis for testing homogeneity is defined assuming different sites have an equal proportion of cases identified as fraud; the alternative hypothesis states that at least one of the sites has a different proportion of fraud. A test of homogeneity is conducted at a significance level of .05 and the null distribution of the usual test statistic is Chi-square with 15 degrees of freedom, since 6 populations of interest (I) and 4 categories (J) are present. The categories involved in the study were the site (Fraud/Nonfraud) and KB (Fraud/Nonfraud).

The proportions involved in the test result from different models trained on particular training sets. A training and a testing set were chosen for each site. The Knowledge Base was divided into two fixed sets with 316 and 116 observations, randomly selected, that are used respectively as training and test sets at all sites.

Table 4.6 summarizes the structure of each site's training and test set. Each of these twelve files consists of a random sample drawn independently from each population site plus the Knowledge Base observations.

Table 4.6 Training and Test set for Analysis of Fraud proportion in different sites.

Site	*Training set*	*Testing sets*
Dayton	*3000 + 316(KB)*	*17560 ; 116 (KB)*
Oakland	*3000 + 316(KB)*	*14500 ; 116 (KB)*
San Antonio [CAPS]	*3000 + 316(KB)*	*17870 ; 116 (KB)*
San Antonio [IAPS]	*3000 + 316(KB)*	*19873 ; 116 (KB)*
San Diego	*3000 + 316(KB)*	*16902 ; 116 (KB)*
Pensacola	*3000 + 316(KB)*	*28400 ; 116 (KB)*

For each site, the least complex model (smallest F tree size) among those with the smallest misclassification rate is selected. Table 4.7 summarizes each site's model selection and respective misclassification risk. The adopted representation of MART models, MART (*M*, *F*) means an ensemble of *M* single *F*-sized trees.

Table 4.7 Selected Models for Analysis of Fraud proportion in different sites.

Site	*Model* *MART(M, F)*	*Test Misclassification risk*
Dayton	*MART(138, 4)*	*0.003*
Oakland	*MART(161, 3)*	*0.000*
San Antonio [CAPS]	*MART(178, 5)*	*0.000*
San Antonio [IAPS]	*MART(118, 4)*	*0.000*
San Diego	*MART(140, 4)*	*0.001*
Pensacola	*MART(145, 4)*	*0.000*

From the result of the analysis we reject the null hypothesis that the proportion of fraud is equal for all the different sites.

It can be seen in Figure 4.11 that significant differences exist between sites, supporting the rejection of homogeneity in site fraud proportions. The “common proportion” line shows the estimated fraud level under the null hypothesis, where all sites’ data can be combined.

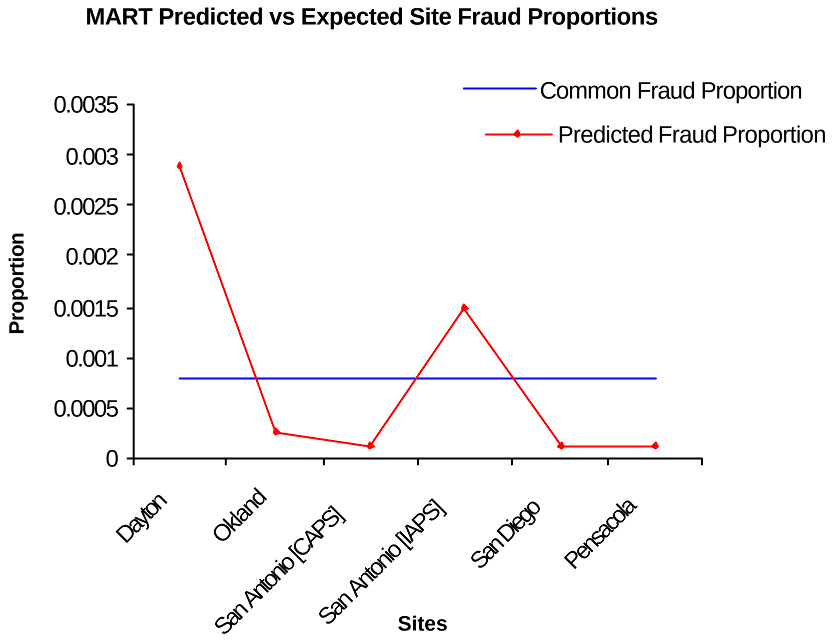


Figure 4.11 MART predicted vs. Expected site fraud proportion.

Multicomparison of the proportions using Bonferoni's inequality show that fraud proportion in Dayton is different at the .05 significance level from all the other sites. San Antonio [IAPS] is also different from all other sites.

Each of these sites, Dayton and San Antonio [IAPS], use the vendor payment system IAPS. This fact can give us an insight into the fact that in the Knowledge Base factors such as technology or a business practice rather than fraud are contributing to train models in predicting fraud.

The other sites have different vendor payment systems, summarized in Table 4.8.

Table 4.8 Sites vendor payment system

Site	*Payment System*
Dayton	*IAPS*
San Antonio [IAPS]	*IAPS*
San Antonio [CAPS]	*CAPS*
Oakland	*STARS*
San Diego	*STARS*
Pensacola	*STARS*

It seems that CAPS payment system is closer in a certain sense to the STARS system than it is to IAPS, which is far way from both CAPS and STARS.

This particular result reveals some insights about the validity of the Knowledge Base to train models for fraud prediction. It is plausible that models trained with the actual fraud historical data are predicting something other than fraud. This drawback and the presence of the missing values strongly suggest that efforts be taken to update the Knowledge Base with fraud cases that use more current accounting systems.

D. CONCLUSIONS

It is possible to identify the relationship and importance that missing values have on fraud classification. The usefulness of partial dependence plots in recognizing this relationship, when derived in parallel with the identification of important variables, identification is clear. In particular, this analysis gives some insights into the way missing values contribute to increasing the odds of observing fraud. Also, conditional dependences can be graphically assessed, which helps to draw conclusions about the predictive relationship between predictor variables and fraud.

The analysis of missing values patterns also reveals asymmetries in fraud data due to a recognized nonrandom pattern.

The MART approach to the problem of missing values was compared to a technique designed to impute missing values, based on prediction probabilities. The unchanged misclassification risk gives support to the way MART deals with and approaches the missing values problem.

The insight that Knowledge Base is training models in classifying and predicting patterns other than fraud constitutes a finding that could contribute to concentrating efforts on improving the actual quality of the Knowledge Base repository.

V. CLASSIFYING FRAUD AND CNI'S

The content of this chapter will focus on analysis of both the Knowledge Base and the CNI database. The set of tools included in MART makes possible the exploration of both databases and offers several insights. The following sections will answer the questions about (a) the identification of relative importance of variables for predicting fraud and CNI classification; (b) the discussion of continuous versus categorical predictor sets for fraud classification; (c) the usage of a binary fraud response variable versus a four-category response variable; and (d) an overview of the MART models' performance at different sites, compared with the results presented in Jenkins [12].

A. IMPORTANT PREDICTORS FOR CLASSIFICATION

The analysis developed in the following subsections offers several insights into fraud prediction information. The identification of the elementary information useful for detecting fraud patterns in the Knowledge Base is presented here through the identification of the relative important variables. This same analysis is then repeated for the CNI database, offering results concerning the identification of the relative important predictors in classifying CNI's.

1. Fraud's Important Predictors

This subsection relies on the identification of the most important variables for fraud classification. The nonfraud transactions mentioned here were selected from the set of CNI4's contained in the CNI database. Training data sets for this analysis were constructed from the Knowledge Base plus a subset of CNI4's (170) from 6 sites audited so far. A set of several MART(*M, F*) models were identified. Each site contributed with a subset of CNI4 added to the known set of fraud transactions, defining the training set. A set of 59 variables, 11 continuous and 48 categorical, constitutes the base for the present analysis. A further analysis about the relative importance of continuous versus categorical variables for fraud classifications is presented in section B.

Figure 5.1 shows the set of important variables for classifying nonfraud. Continuous predictors' names are followed by (*). The horizontal scale is weighted one in which importance is computed based on a voting scheme that measures the relative importance of each variable present on each site's MART(*M,F)* model. Here, the weighted relative importance of each predictor is the result of the sum of the individual measure of importance, on each model, divided by the largest sum of importance between all selected predictors. The final result is a weighted relative importance rescaled from 0 to 1, with 1 being assigned to the variable with the largest relative importance.

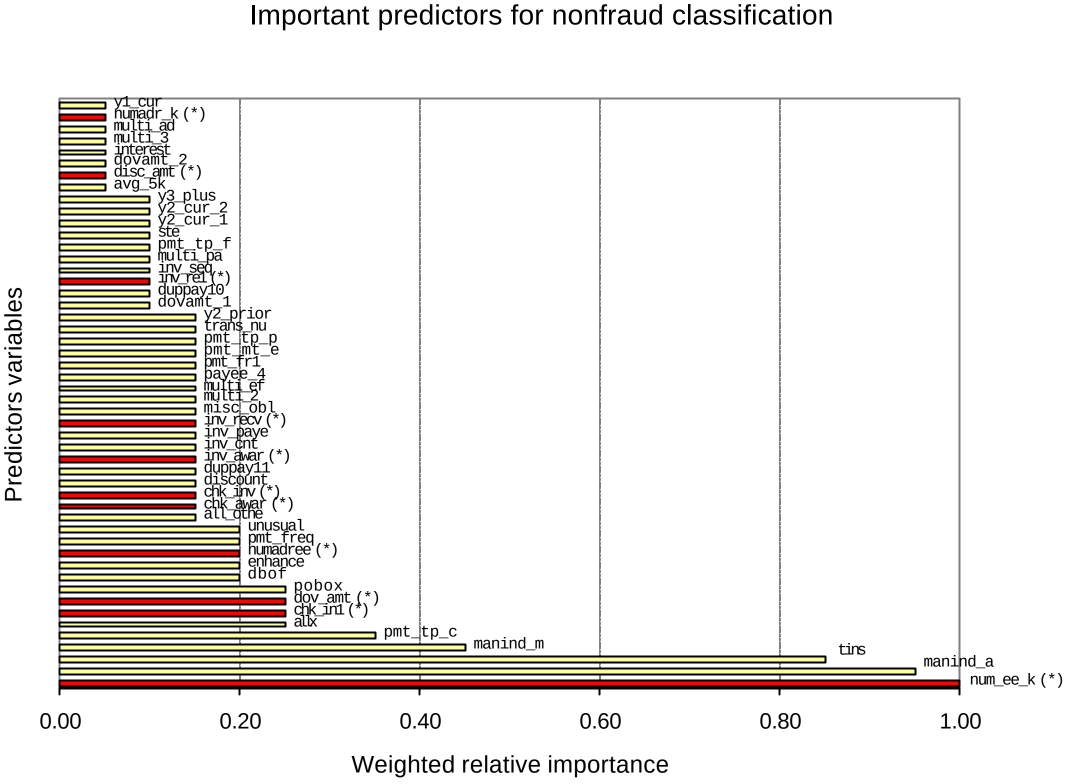


Figure 5.1 Variable Importance for classifying nonfraud

The process of identification of relative important variables for classifying nonfraud did not included nine of the categorical variables. The set of variables with no relative importance, 'BRAC', 'FEW_PYMT', 'MILPAY', 'MULTI_TINS_K', 'MULTI_TINS', 'NOT_DFAR', 'OTHERX', 'PMT_METH_D', 'Y1_PRIOR', identifies the variables never present in any ensemble defined for nonfraud classification.

Figure 5.2 shows the set of relative important variables for classifying Bigsys.

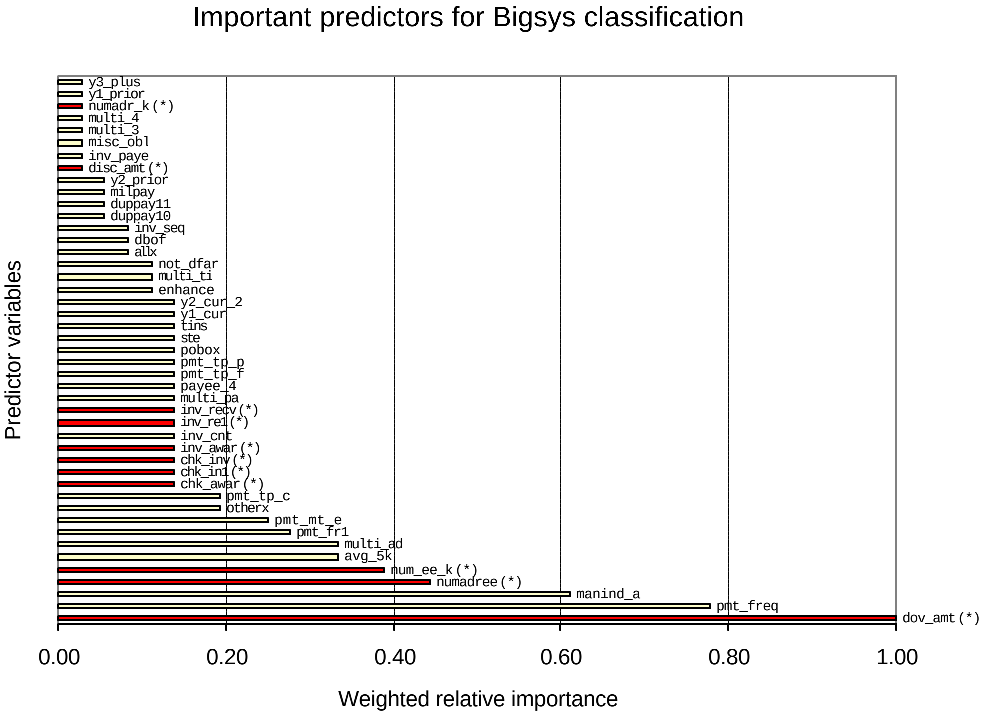


Figure 5.2 Variable Importance for classifying Bigsys

For classifying Bigsys from an initial set of 59 candidate predictor variables, the following set of 14 categorical variables show no importance: ‘ALL_OTHER’, ‘BRAC’, ‘DISCOUNT’, ‘DOVAMT_1K’, DOVAMT_2K’, ‘FEW_PYMT’, ‘INTEREST’, ‘MANIND_M’, ‘MULTI_PAYE’, ‘MULTI_EFT_K’, ‘PMT_METH_D’, ‘TRANS_NUM’, ‘UNUSUAL’, ‘Y2_CUR_1ST’. Results were reviewed with the data mining staff for reasonableness, revealing that the selected predictors made sense in predicting Bigsys; the top 3 predictors in order of relative importance are the continuous variable “DOV_AMT” (the amount of the payment) and the categorical variables “PMT_FREQ” (a flag indicating the frequency of payments to the vendor) and “MANIND_A” (a flag indicating whether the payment was automatic or not).

Figure 5.3 shows the relative important variables for classifying Opportunistic fraud category.

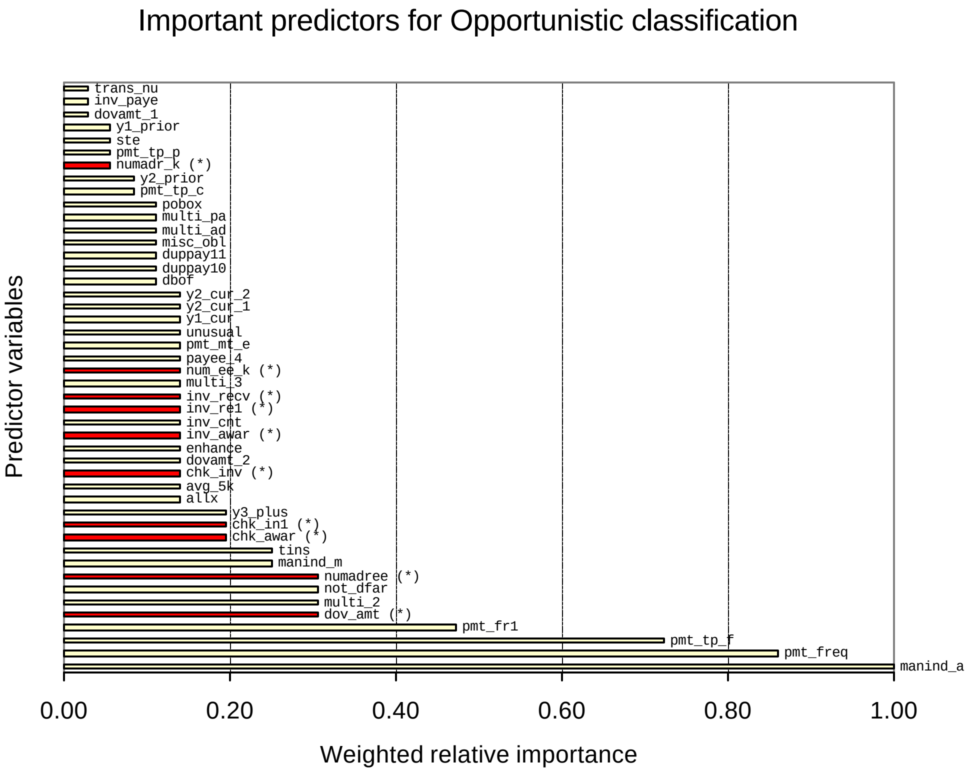


Figure 5.3 Variable Importance for classifying Opportunistic

The set of variables not contributing for classifying Opportunistic fraud category includes 12 categorical variables 'ALL_OTHER', 'BRAC', 'DISCOUNT', 'FEW_PYMT', 'INTEREST', 'INV_SEQ', 'MILPAY', 'MULTI_TINS_K', 'MULTI_EFT_K', "MULTI_TINS', 'OTHERX', 'PMT.METH.D' and the continuous variable 'DISC_AMT'. The top 3 predictors in order of relative importance are the categorical variables "PMT_FREQ" (a flag indicating the frequency of payments to the vendor), "MANIND_A" (a flag indicating whether the payment was automatic or not) and "PMT_TP_F" (a flag indicating whether or not the payment was a final or partial payment).

Figure 5.4 shows the relative important variables for classifying Piggy fraud category.

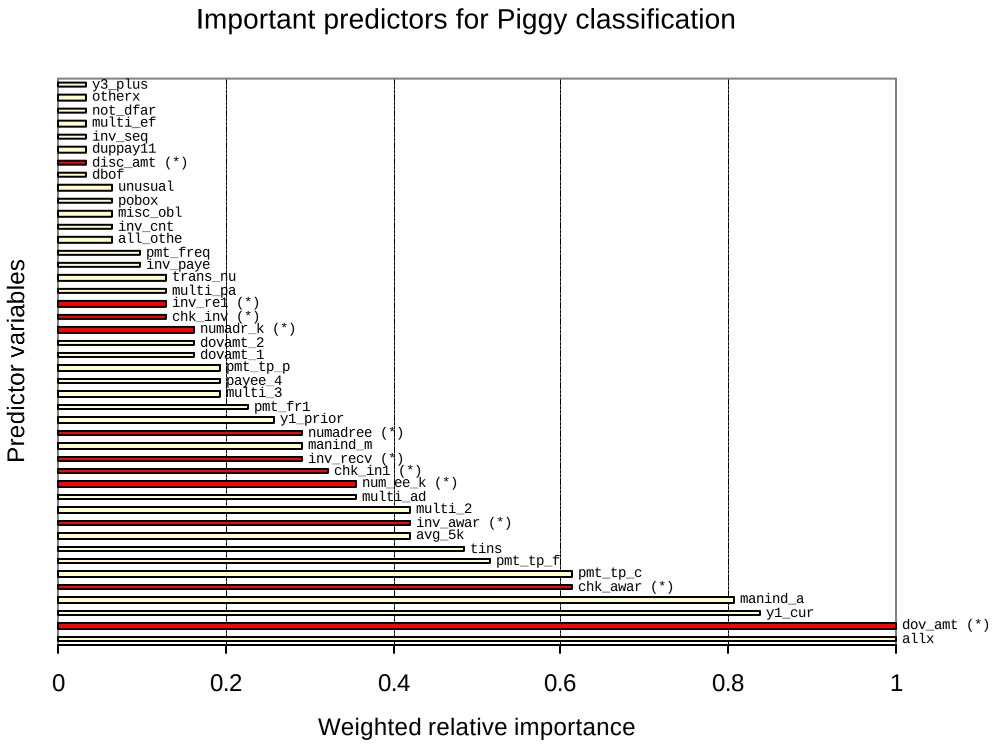


Figure 5.4 Variable Importance for classifying Piggy

In Piggy classification the set of candidate predictors not contributing to the classification includes the following 15 categorical variables ‘BRAC’, ‘DISCOUNT’, ‘DUPPAY102’, ‘ENHANCE’, ‘FEW_PYMT’, ‘INTEREST’, ‘MILPAY’, ‘MULTI_TINS_K’, ‘MULTI_TINS’, ‘PMT_METH_D’, ‘PMT_METH_E’, ‘STE’, ‘Y2_CUR_1ST’, ‘Y2_CUR_2ND’ and ‘Y2_PRIOR’. The top 3 predictors in order of relative importance are the categorical variables “ALLX” (a flag indicating whether or not the payment was made from an “X” year appropriation) and “Y1_CUR” (a flag indicating whether or not the payment was made from a ‘1 year current’ appropriation), and the continuous variable “DOV_AMT” (the amount of the payment).

Figure 5.5 shows the relative important variables for classifying Smallsys fraud category.

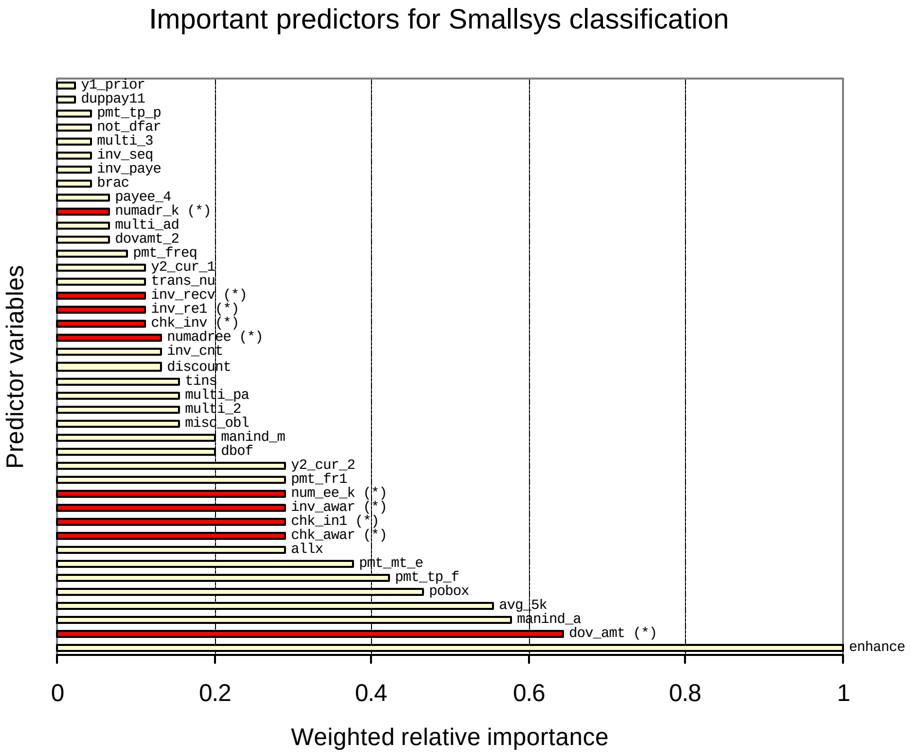


Figure 5.5 Variable Importance for classifying Smallsys

For Smallsys classification the set of candidate predictors not contributing for its classification includes 17 categorical variables ‘ALL_OTHER’, ‘DOVAMT_1K’, ‘DUPPAY102’, ‘FEW_PYMT’, ‘INTEREST’, ‘MILPAY’, ‘MULTI_TINS_K’, ‘MULTI_EFT_K‘, ‘MULTI_TINS’, ‘OTHERX’, ‘PMT_METH_D’, ‘PMT_TYPE_C’, ‘STE’, ‘UNUSUAL’, ‘Y1_CUR’, ‘Y2_PRIOR’, ‘Y2_PLUS’ and the continuous variable ‘DISC_AMT’. The top 3 weighted relative important variables for classifying Smallsys are the categorical variables “ENHANCE” (a flag indicating whether or not the first address line is populated with the payee) and “MANIND_A” (a flag indicating whether the payment was automatic or not) and the continuous variable “DOV_AMT” (the amount of the payment).

In the overall process of fraud/ nonfraud classification it was found that variables 'FEW_PYMT' and 'PMT_METH_D' have never been involved in classifying any category of fraud or nonfraud.

2. CNI's Important Predictors

This subsection deals with the identification of important variables for CNI's identification. It is interesting to compare the results for nonfraud important variables with CNI4 classification important variables. The nature of both nonfraud and CNI4 is the same, but the presence of distinct patterns associated with fraud cases and those related to CNI's determine the importance of each predictor for classifying each pattern.

This analysis identified two variables that had never been involved in the process of splitting the different CNI categories, 'MANIND_A' and 'PMT_METH_D'.

The representative plots of the most important variables in classifying Serious CNI's (CNI1), CNI's (CNI2) and noCNI (CNI4) are shown in Appendix E.

The fact that different variables are important in classifying CNI's and Fraud does not by itself indicate that CNI and Fraud are different types of cases. However, this together with the observation that the models for which fraud prediction classification rate are high, tend to do more poorly in predicting CNI's than models with less ability to predict fraud, indicate that CNI's and Fraud records may not be as related as previously thought [12]. One potential source of difference is that auditors classify CNI's where Fraud cases have been prosecuted. The alternate theory (especially for Bigsys and Smallsys) is that people who are committing fraud might go through extra lengths to ensure there are no CNI's on the fraudulent voucher, which would call attention by the audit staff. It may be that modeling on CNI's is a great way to predict CNI's in a population, which is useful to the audit staff, but not a good way to predict fraud in a population. DFAS efforts are now underway to look more closely at the consistency of audits across sites.

B. TRAINING SETS

This section addresses two main issues: the first, about the practice of categorization continuous variables into categorical variables and the second about the best response variable, fraud-4cat (Bigsys, Opportunistic, Piggy, Smallsys and nonfraud) versus fraud-binary (fraud/nonfraud) for supporting model training. The amount of time spent in categorizing continuous variables, reflects the arduous task DFAS has in trying to define or identify thresholds to create these categorical variables. The potential benefit that categorical variables are handled more easily and contribute models running faster needs to be balanced against the fact that appropriate thresholds might evolve over time and might differ from site to site. Example of this is the dollar amounts associated with contracts, which are subject to fiscal and inflation changes. In addition, changes in accounting procedures will shift thresholds, requiring a new analysis and redefinition. The way MART deals with continuous and categorical variables offers an opportunity to explore this problem and observe the results MART selects for identifying important variables for fraud classification. In addition, easily dealing with continuous and categorical variables, MART is also faster than CART and NN, and easier to implement.

1. Continuous versus Categorical Candidate Predictors

This sub-section gives an overview about the performance of MART models trained on different training sets. It includes models with continuous and categorical variables (CaCo), models with only categorical (Ca) variables and models with only continuous variables (Co).

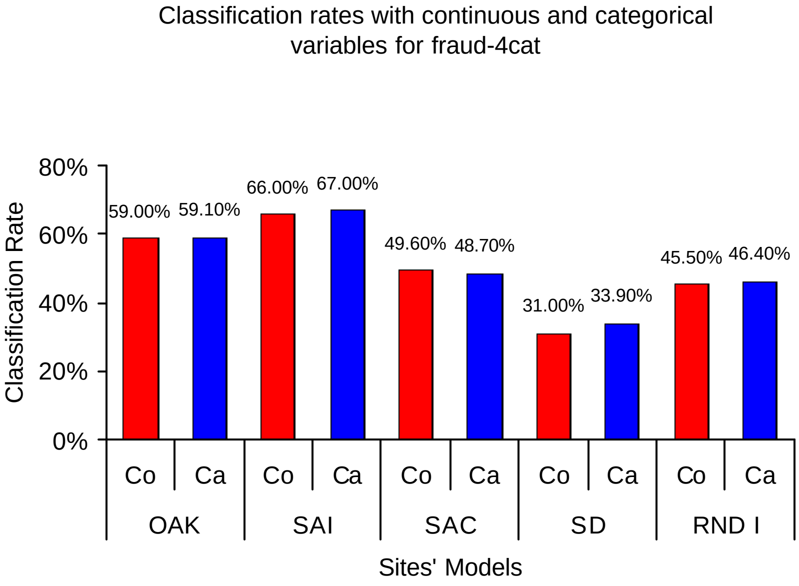


Figure 5.6 Classification rates for models with continuous and categorical variables

Figure 5.6 presents a performance overview of different models trained on training sets with continuous (Co) and categorical (Ca) variables. With the exception of SD, there is not real benefit to using the categorical variables over the continuous variables. The fact that the categorical variables perform about the same as the continuous variables supports the idea that DFAS is identifying appropriate thresholds for each continuous variable.

Figure 5.7 gives the classification rates for site models trained on continuous (Co), categorical variables (Ca) and the combination of both continuous and categorical (CaCo) variables with a fraud-4cat response. In particular, it gives an odd pattern of CaCo classification rates associated with OAK and SD sites. Figure 5.8 gives a slightly different view; it includes an additional case, site models trained with fraud as a binary response. This plot shows the dispersion of classification rates for each set of candidate predictors from the Knowledge Base, using MART models at different sites.

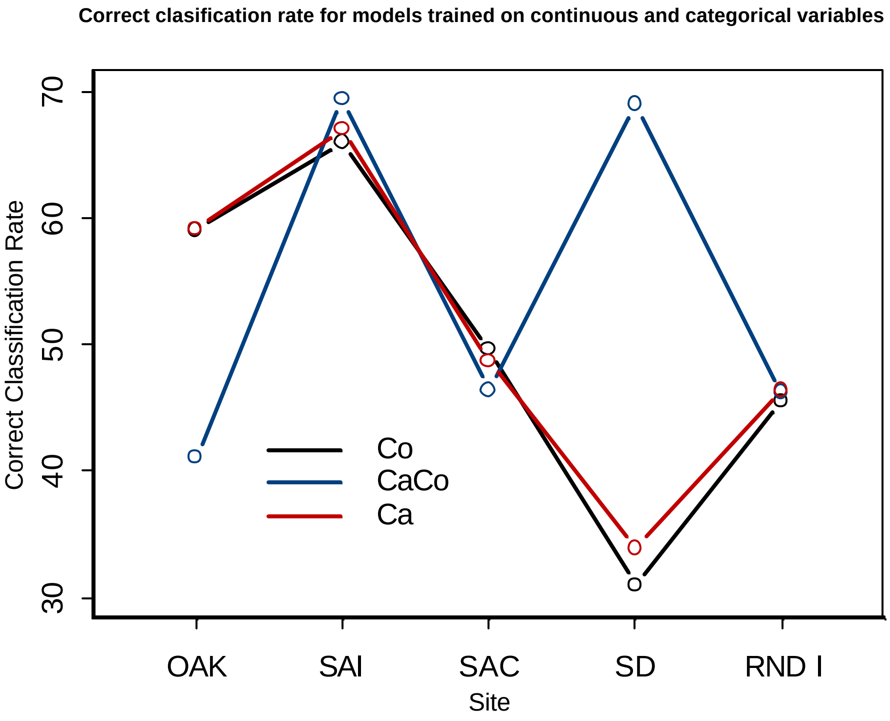


Figure 5.7 Correct classification rate for models trained on continuous and categorical variables for fraud-4cat

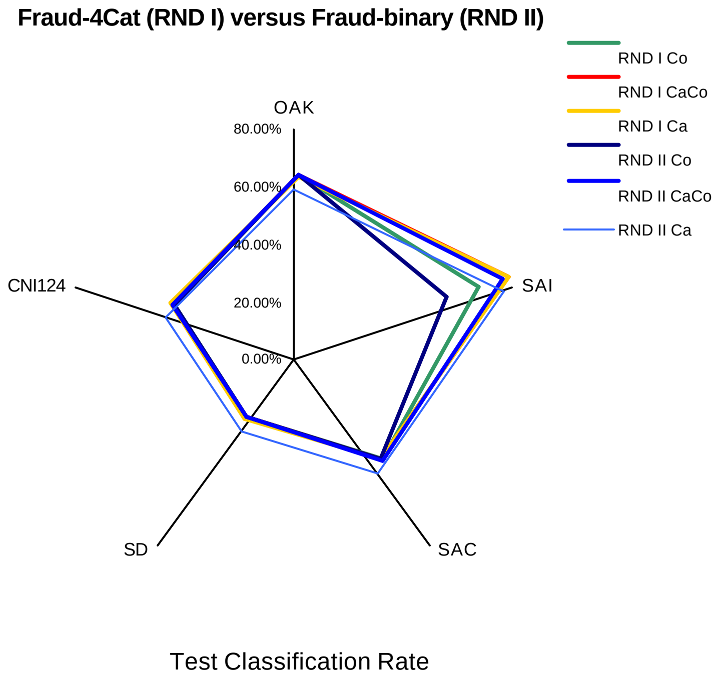


Figure 5.8 Test classification rates for MART models trained with binary fraud and four-fraud category

The analysis with continuous variables (RND I/II Co), continuous and categorical variables (RND I/II CaCo) and only categorical variables (RND I/II Ca) show that all sites but SAI presents similar classification rates for different training set configurations. A more formal MANOVA analysis shown that no significant differences exist in the mean classifications rates between models where only original continuous variables are used, models where only the derived categorical variables are used and models where both the original and derived variables are used.

2. Fraud Response Variable

A question for Mongoose staff was whether the codification of fraud as binary (RND II) rather than as four-fraud category (RND I) would improve the ability to predict fraud. We expect with MART that classifying fraud as binary will give results comparable to classifying fraud with more categories (fraud-4cat) because of the method MART (and sometimes classification trees) uses to fit models with categorical responses that have more than one level. In order to clarify this question, it is demonstrated that no enhancement is achieved by treating fraud as binary.

Figure 5.8 shows that classification rates do not change significantly when MART(*M*, *F*) models have to predict fraud classified as binary (RND II) instead of with four fraud categories (RND I). The training set for RND I/ II MART models is based on the Knowledge Base plus a set of CNI4 randomly selected from the 6 sites so far audited. The radial plot presents level curves, showing that classification rates assigned to model RND I (four fraud categories; Bigsys=1, Opportunistic=2, Piggy=3, Smallsys=4, nonfraud=0) and RND II (binary fraud; fraud=1, nonfraud=0) in four different sites and over the CNI database do not change significantly. Different fraud categorization do not show influence in MART(*M*, *F*) models classification rates for different sets of data (OAK, SAC, SD, CNI124) and predictor variables (Co, CaCo, Ca). Only the SAI site shows some difference in classification rates.

Figure 5.9 and Figure 5.10 show respectively the performance of MART models trained on a four-fraud category response variable and a binary fraud response variable. In Figure 5.9, for the four-fraud category response variable, models with categorical variables performs equally well for all sites except SAI (almost perfect equilateral triangles).

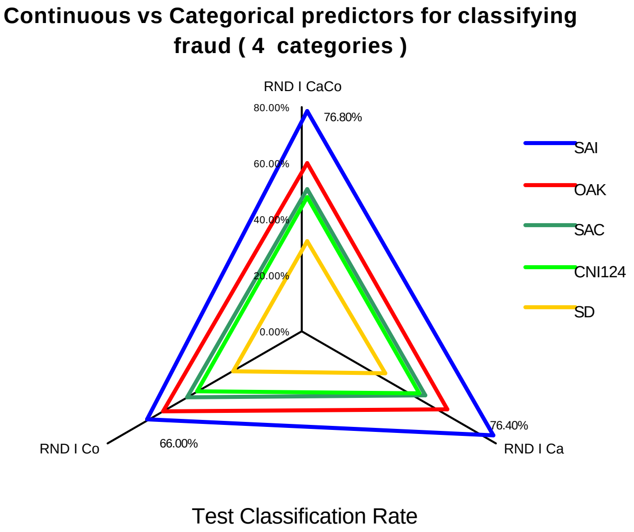


Figure 5.9 Test classification rates for MART models trained on four-fraud category

Figure 5.10 shows that, for a binary fraud response variable, identical behavior is observed for the different sites. For SAI the classification rate for the MART model trained on continuous predictors with a fraud-binary response is substantially reduced.

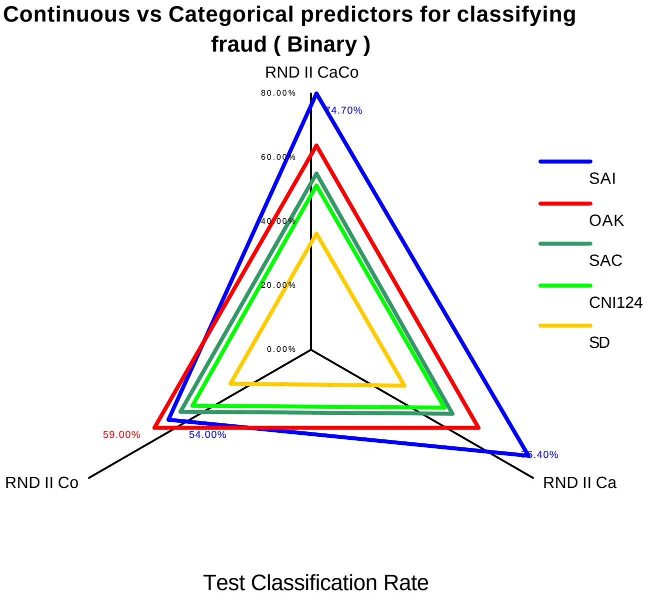


Figure 5.10 Test classification rates for MART models trained on binary fraud category

We note that neural networks and some implementations of CART will give different results when fraud is classified as binary (fraud/nonfraud) from when it is classified in four categories. MART however does not. Because of the NN design, it is expected computational advantages of using response variable fraud as binary (fraud/nonfraud).

3. Importance of Continuous Predictors

The analysis included eleven continuous variables ('CHK_AWARD_DT', 'CHK_INV_DT', 'CHK_INV_RECV_DT', 'DISC_AMT', 'DOV_AMT', 'INV_AWAR', 'INV_RECV_IND_DT', 'INV_RECV_AWARD_DT', 'NUM_EE_K', 'NUMADR_K', 'NUMADREE'); however ten show importance in classifying fraud. Variable 'DISC.AMT' was never present in classifying fraud. The DFAS data mining staff revealed that 'DISC.AMT' is not associated with fraud because makes the payment more visible to audit staff.

Figure 5.11 shows the most important continuous variables for classifying nonfraud. The variables 'DISC.AMT' and 'CHK_AWARD_DT' show no importance in classifying nonfraud from fraud categories.

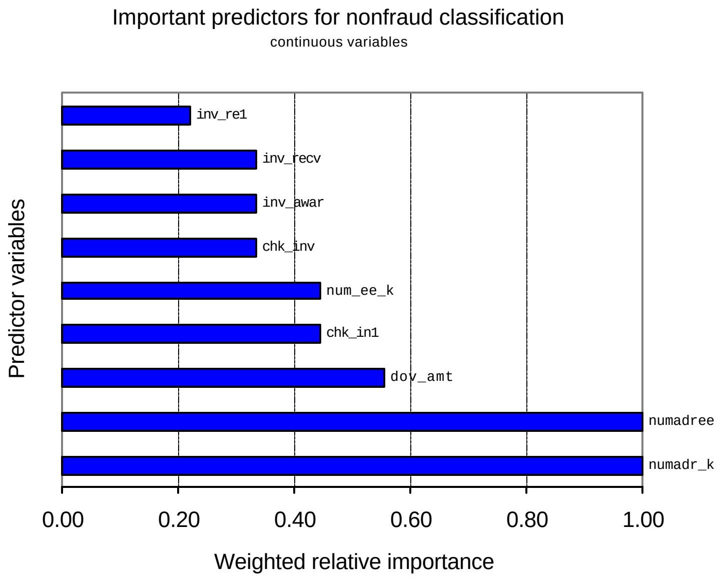


Figure 5.11 Variable importance for classifying nonfraud with continuous variables

Figure 5.12 to Figure 5.15 give the relative important predictors for each fraud category, respectively Bigsys, Opportunistic, Piggy and Smallsys.

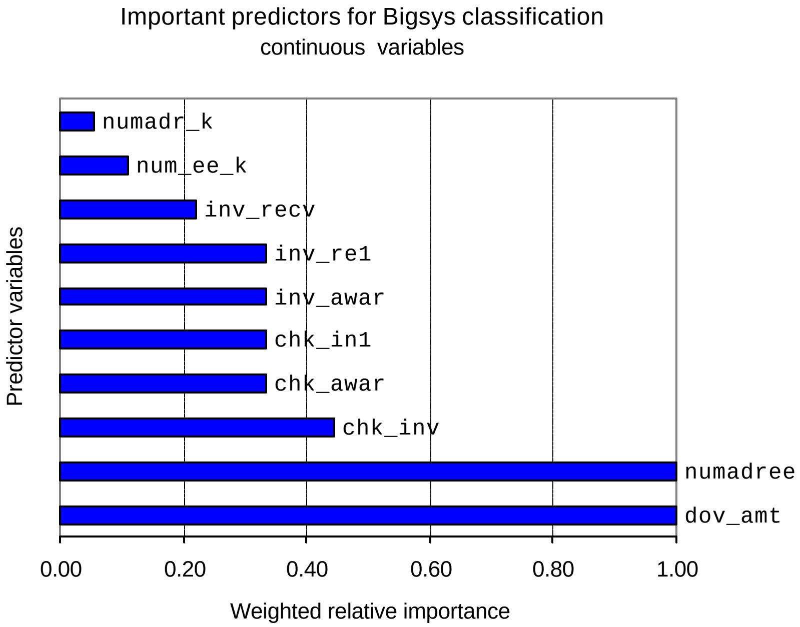


Figure 5.12 Variable importance for classifying Bigsys with continuous variables

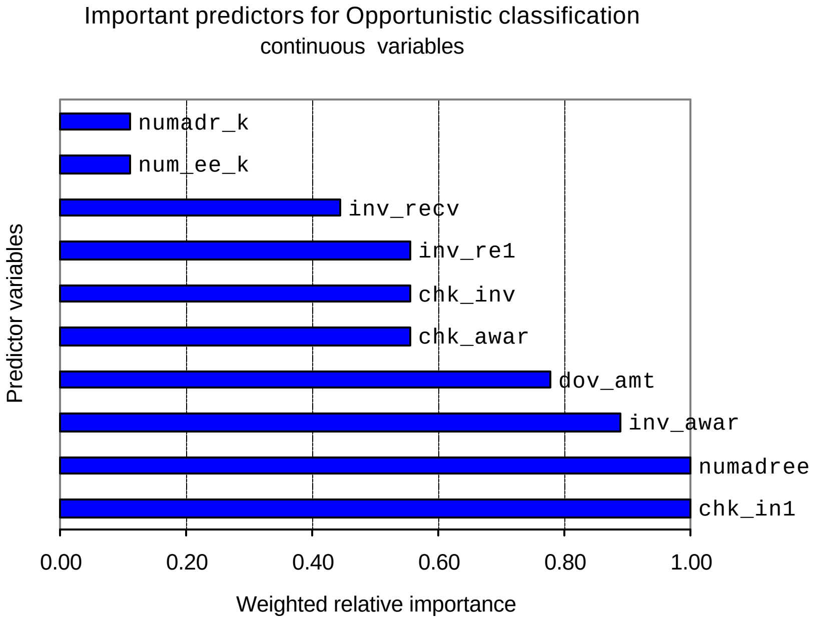


Figure 5.13 Variable importance for classifying Opportunistic with continuous variables

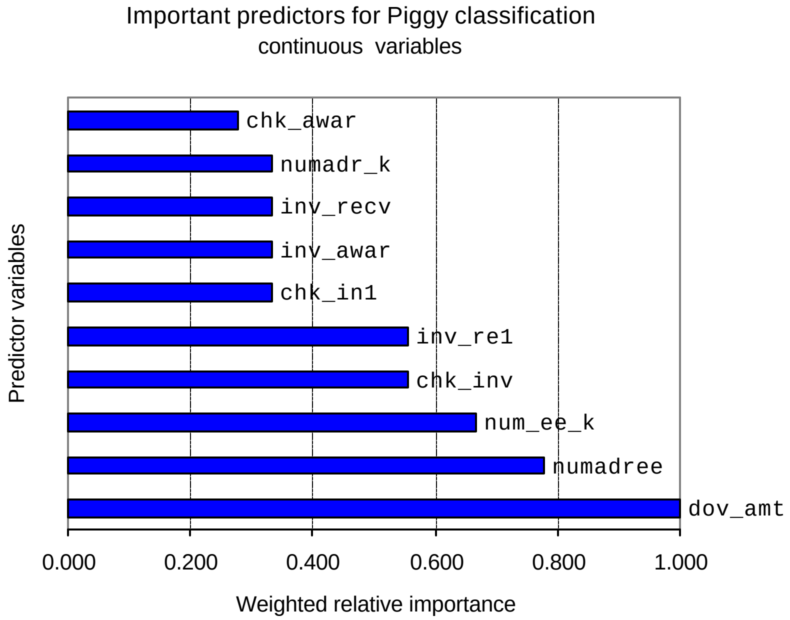


Figure 5.14 Variable importance for classifying Piggy with continuous variables

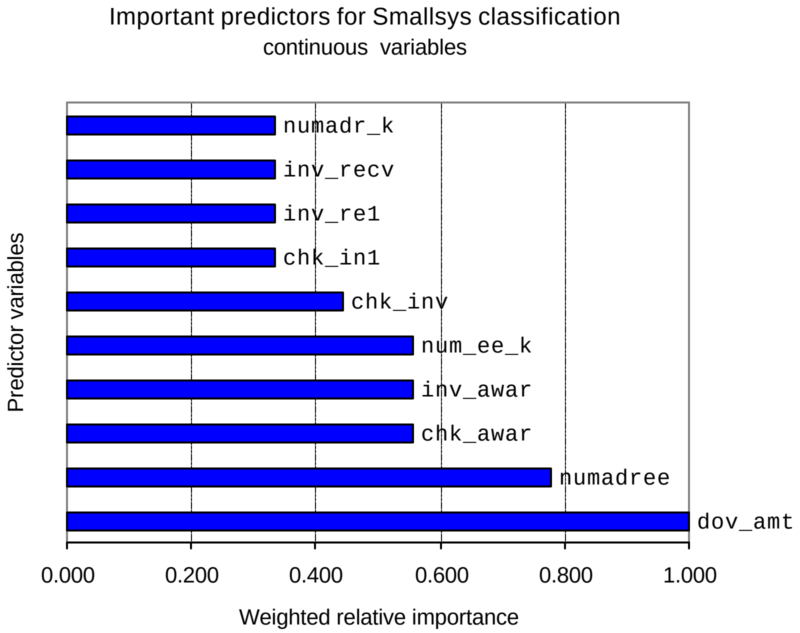


Figure 5.15 Variable importance for classifying Smallsys with continuous variables

C. MART MODEL PERFORMANCE

This section compares the performance of the different models (C5, NN) developed by DFAS (Appendix D gives the four sites' model names and CNI classification rates [12]) and the following MART models trained on the Knowledge Base plus: (1) MART operating on a subset of the CNI 4 (nonfraud) from each particular site (SD, OAK, SAC, SAI); (2) MART.R operating on a subset randomly selected from all CNI4 from the 6 sites so far audited.

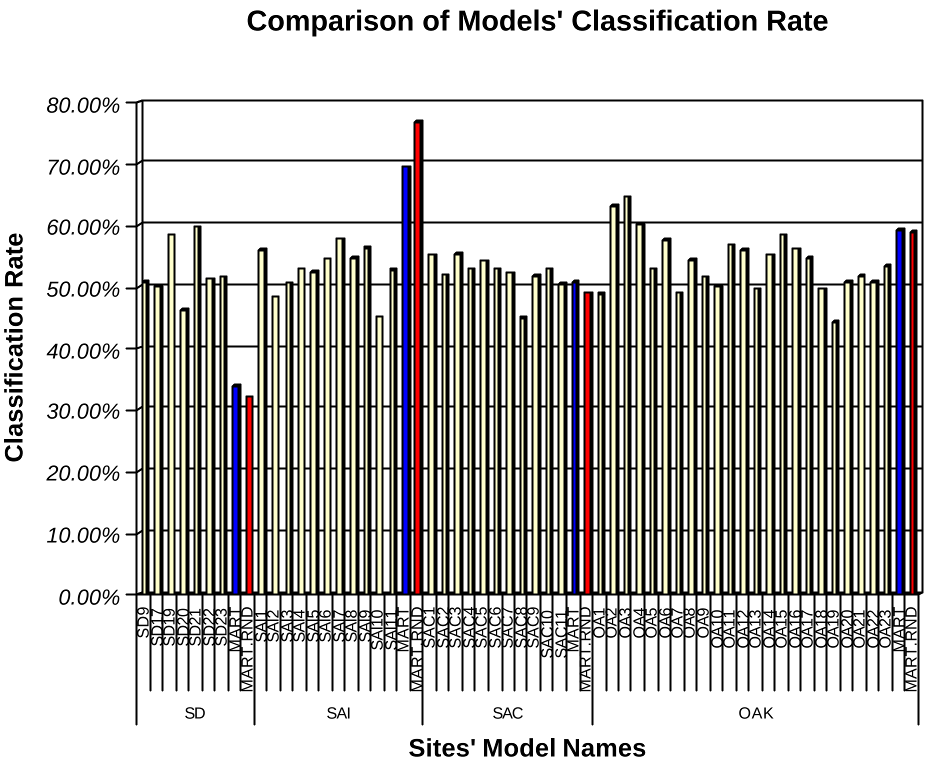


Figure 5.16 Comparison of Models' classification rate

The analysis of the Figure 5.16 gives some insights about the influence of the CNI4's origin in the classification rate. The same MART(*M, F*) model, MART and MART.R, give slightly different classification rates for different sites. Only in SAI is the difference between the two models substantial (7.4% in SAI versus SD 1.9%, SAC 1.7%, OAK 0.2%). Also, in SAI the MART methodology greatly outperforms models based on

C5 or NN. The performance of MART models in OAK and SAC is close to the C5 and NN models and in SD MART is worst at predicting CNI's. Different MART(*M*, *F*) models trained with different subsets of CNI4's produce classification rates that are quite different from the other models in SAI and SD but only slightly different on OAK, SAC and CNI124 (random test set from CNI database with 6 sites).

D. CONCLUSIONS

The identification of the most important variables for fraud and CNI classification is produced as the result of a weighted vote of different MART(*M*, *F*) models. In particular the set of important variables for classifying nonfraud and CNI4 have significant differences, expressing in a certain sense the distance between fraud and CNI patterns. For a threshold of 80% relative importance in fraud classification, the variables 'NUM_EE_K', 'MANIND_A' and 'TINS' with relative importance of 1.0, 0.95 and 0.85, give a relative importance in CNI classification of 0.16, 0.30 and 0.00 respectively. The same effect is seen in the opposite direction with 'DOV_AMT' (1.00), 'CHK_AWAR_DT' (0.99), 'INV_RECV_IND_DT' (0.87) and 'CHK_INV_DT' (0.85) that give relative importance in fraud classification of 0.25, 0.15, 0.10 and 0.15 respectively. Thus, the intent of seeking fraud using the CNI database for training requires additional analysis. A preliminary study, conducted over the present fraud and CNI databases, suggested that as fraud prediction classification rate increase, the ability to predict CNI's decreases.

In addition of the question of converting continuous variables into categorical predictors has been addressed. An overview of the performance of different models trained on different variables' training sets, including models with both continuous and categorical variables, models with only categorical variables and models with only continuous variables has been presented. The performance of MART(*M*, *F*) models trained on sets of continuous variables gave a performance close to that the MART(*M, F*) models trained on sets of categorical variables. The conclusion is time can be saved by using continuous variables directly in MART models.

In the overall process of fraud/ nonfraud classification it was found that variables 'FEW_PYMT' and 'PMT_METH_D' have never been involved in classifying any category of fraud or nonfraud. Also, when continuous variables by itself are training models, 'DISC_AMT' show no contribution to fraud classification.

The analysis of the advantage of using a fraud-binary response variable versus a fraud-4cat response variable, showed, as expected, no significant differences for MART models; also, for CART models we expect identical performance for models trained on fraud-binary and fraud-4cat response variable. Because of the NN design, we expect that using response variable fraud as binary (fraud/nonfraud) will be computationally advantageous.

The comparison of performances of different models (C5, NN) so far developed by DFAS and the MART models trained on the Knowledge Base plus: the MART and MART.R described earlier exposed some insights about the influence of the CNI4's origin (site) in the classification rate. The same MART(*M*, *F*) model, MART and MART.R, give slightly different classification rates between them for different sites. Thus, the idea of seeking fraud using the CNI database for training requires additional analysis. Other considerations about the nature of the present CNI information should be explored, such as the auditors' constancy of procedure in inspecting transactions, in order to avoid biased patterns from being present in the CNI database. This is required to pursue the intent of fraud detection based on CNI analysis, as supported so far in the literature ([12].)

VI. CONCLUSIONS AND RECOMENDATIONS

A. CONCLUSIONS

The study of the applicability of the MART methodology to DFAS fraud detection usage, focuses on four data related issues: (1) the Knowledge Base study that includes a missing values analysis; (2) the identification of the important variables for fraud and CNI classification; (3) the study of models trained on continuous versus categorical variables; (4) a comparison of MART models' performance versus C5 and NN models. The most relevant results concerning the applicability of this new methodology is its ability to deliver results within a few hours comparable to or better than those requiring months of hands-on development by expert data mining teams. In that sense, MART should be seen as a new methodology in the DFAS data mining process, assuring a faster data knowledge progress that might improve fraud detection ability.

1. A New Methodology

MART methodology is shown to be an alternative tool for improving the current process of predicting fraud and CNI's. This methodology should be seen in an integrated knowledge environment where additional information and process improvements are offered. Advantages of introducing MART in the DFAS fraud detection process include the facts that (1) it is not very sensitive to data errors or outliers in predictors or the target variable; (2) needs no time-consuming binning activities; (3) permits the selection of different type (continuous and categorical) candidate predictors without any previous data preparation (data does not require transformation, or other time-consuming processing); (4) missing values are handled automatically, contributing to the identification of dependences in the classification process; (5) it is resistant to over-training and models reach their maximum accuracy well before 1,000 trees are grown and can be effectively trained with only about 20% of the data; and (6) presents a high speed of model development.

2. Training Models With Knowledge Base versus CNI Database

The insight that the Knowledge Base is potentially training models to classifying and predicting patterns other than fraud contributes to arguments that the Knowledge Base repository should be updated as current fraud cases became available. This will help identify changes or mutations in fraud patterns motivated by fraud perpetrators' intelligence as well as by technology evolution or new process transactions. The problem of an old Knowledge Base, not covering current business practices such as EFT payments, and having problems identified here such as missing values is addressed in the present study. Research on the application of the MART methodology to the CNI database is needed.

The Norfolk site (August 2002) was the primary one to be studied using the MART methodology for training models on the CNI Database. The purpose was to predict CNI's present in about 22,775 CAPS transactions. Based on a voting system of different MART models, thirty-one transactions out of 97 predicted as Serious-CNI or CNI were selected by the Mongoose team, and included as audit referrals for the site visit. The results of that inspection will be presented in the site report.

3. About Missing Values

The missing values analysis included in the Knowledge Base study revealed the relationship and importance missing values have on fraud classification. The usefulness partial dependence plots have in recognizing those patterns, when derived in parallel with identification of important variables, becomes a major issue in this study for better understanding fraud patterns present in the Knowledge Base. In particular this research gives some insights about the way missing values contribute for increasing the odds of observing fraud.

The analysis of missing values pattern also reveals asymmetries in fraud data. A recognized nonrandom pattern of missing values might contribute to difficulties in fraud prediction.

A study of imputing missing values support the way MART methodology handles the missing values problem and reveals that no fraud prediction advantage is offered by imputing values on missing valued predictors present in the actual Knowledge Base.

4. Identifying Important Variables

The identification of relative importance of variables for classifying fraud and CNI's highlights the most relevant predictors present in the Knowledge Base and CNI database. The most important variables for fraud and CNI classification have been presented as the result of a weighted vote of different MART(*M, F*) models. In particular the set of important variables for classifying nonfraud and CNI4 present significant differences expressing in a certain sense the distance between fraud and CNI patterns. For a threshold of 80% relative importance in fraud classification, the variables 'NUM_EE_K', 'MANIND_A' and 'TINS' with relative importance of 1.0, 0.95 and 0.85, present a relative importance in CNI classification of 0.16, 0.30 and 0.00 respectively! The same effect is registered in the opposite direction with 'DOV_AMT' (1.00), 'CHK_AWAR_DT' (0.99), 'INV_RECV_IND_DT' (0.87) and 'CHK_INV_DT' (0.85) that present relative importance in fraud classification of 0.25, 0.15, 0.10 and 0.15 respectively. Thus, the idea of seeking fraud using the CNI database for training requires additional analysis. A preliminary study, conducted over the present fraud and CNI databases, suggested that as fraud prediction classification rate increase, the ability to predict CNI's decrease.

5. Continuous versus Categorical Variables

The study of MART model performance when trained on continuous versus categorical variables for predicting fraud supports the fact that data do not require being transformed, or preprocessed in any way, for MART training purposes. This fact reveals a major advantage DFAS can explore using this methodology, saving time used to convert continuous variables into categorical ones. MART(*M, F*) models trained on sets of continuous variables performed about as well as the MART(*M, F*) models trained on categorical set of variables.

For the purposes of supporting other data mining methodologies, the argument that categorical variables are handled easily and contribute to models running faster has to be balanced against the fact that the ideal cut points might evolve and differ from site to site.

6. Binary versus 4-Category fraud

The analysis of the advantage of using a fraud-binary response variable versus a fraud-4-category response variable, show, as expected, no significant differences for MART models.

7. MART versus C5 & NN

The comparison of performances of different models (C5, NN) so far developed by DFAS and the MART models trained on the Knowledge Base plus: (1) MART - a subset of the CNI 4 (nonfraud) from each particular site (SD, OAK, SAC, SAI); (2) MART.R - a subset randomly selected from all CNI4 from the 6 sites so far audited, exposed some insights about the influence of the CNI4's origin (site) in the classification rate. The same MART(*M*, *F*) model, MART and MART.R, produces slightly different for different sites. Other considerations should be explored about the nature of the present CNI information, such as the auditors' constancy of procedure in inspecting transactions, in order to avoid biased patterns being present in the CNI database. This is required to pursue the intent of fraud detection based on CNI analysis, as supported so far in the literature, as reported in Jenkins [12].

B. RECOMMENDATIONS

Promote the inclusion of MART methodology in the DFAS data mining process. MART automatically handles missing values, allows an automatic selection of candidate predictors without preprocessing, it is resistant to over-training and it is fast.

Promote the inclusion of continuous predictor variables for training models with MART methodology, which do not requires time-consuming operations of binning on transformation.

Promote the update of the Knowledge Base, as a primary goal to support fraud detection.

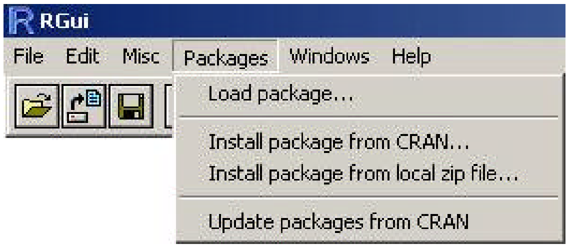
Promote the analysis of the CNI database, regarding the exploration of biased patterns due to auditors' non-constancy of procedures, and also train new models directly on CNI information, to be improved after each new site is audited. The identification of CNI's supports fraud detection in the sense that reducing CNI's contributes to a controlled environment resistant to the perpetration of fraud.

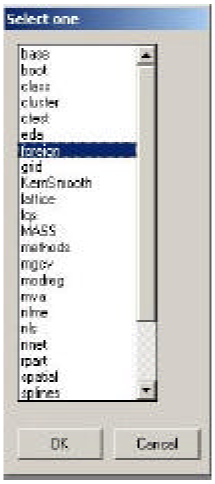
THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A EXPLORING MART IN R – COMMAND REFERENCE

Importing Data from SPSS

In order to import data from the SPSS (files *.sav) load the R package *foreign:*


RGui
File
Edit
Misc
Packages
Windows
Help
Load package...
Install package from CRAN...
Install package from local zip file...
Update packages from CRAN



```
> oak.kbcni <- read.spss('G:/KBxCNI/kb_oakcni.sav')
```

Formatting Data inside R

Selecting column 60 (response variable) from oak.kbcni and assign it as a new object oak.kbcni.y:

```
> oak.kbcni.y <- oak.kbcni[60]
```

Selecting columns 1 to 59 (candidate predictor variables) from oak.kbcni and assign it as a new object oak.kbcni.x:

```
> oak.kbcni.x <- oak.kbcni[1:59]
```

MART command *mart(pred, resp, …)* requires that *pred* and *resp* be matrix objects:

```
> oak.kbcni.y <- as.matrix((oak.kbcni.y))
> oak.kbcni.x <- as.matrix((oak.kbcni.x))
```

The following command defines a vector that identifies which variables from the predictor set (oak.kbcni.x) are numerical (1), categorical (2) or are excluded (0) from the training operation:

```
> kbcni.lx <- c(rep(2,9), rep(1,40), rep(2,3), rep(1,7))
```

Exploring MART inside R

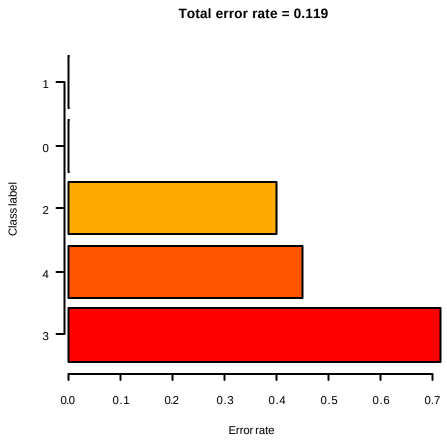
The command *mart(pred, resp, vector, martmode=’class’, tree.size=#)* starts the MART model definition, specifying the parameters *martmode* and *tree.size*:

```
> mart(oak.kbcni.x, oak.kbcni.y,kbcni.lx,martmode='class',tree.size=3, cat.store=1500000)
MART execution finished.
  iters      best   test misclass risk
   70         33      0.1192
```

Analyzing the plots *progress()* we might request MART to go further using *moremart()* to find an ensemble with smallest test misclassification risk:

```
> moremart()
 MART execution finished.
  iters      best   test misclass risk
   270         33      0.1192
```

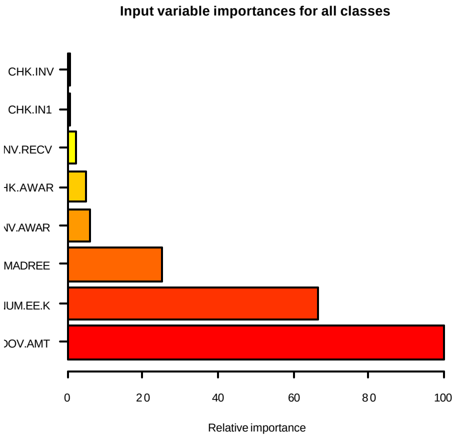
```
> classerrors()
```



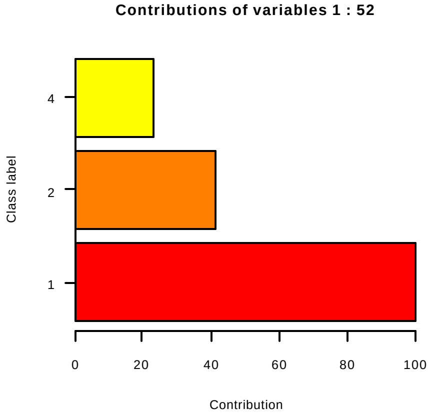
Note: All the classes are ordered in decreasing value of their error rate. The plot represents a graphical examination of the class error matrix for the test data set.

```
> varimp(range=1:8)
```

Note: Plot of the relative predictive importance of predictor variables of the current model. The length of each bar is proportional to the estimated relevance of the correspondingly labeled input variable in making model predictions. The variables are plotted in sorted order of relevance.

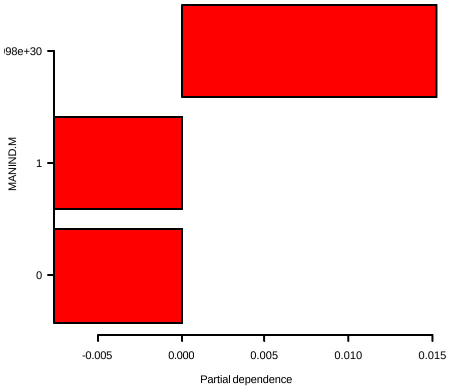


> classimp()



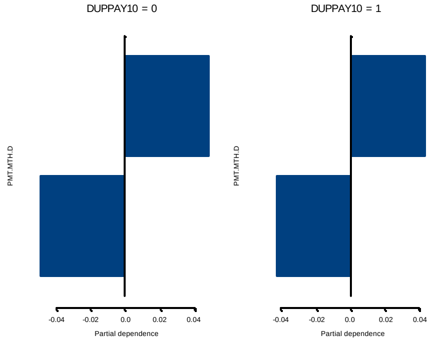
Note: Plot which classes benefit most from the presence of particular set of predictor variables. The length of each bar is proportional to the estimated benefit the correspondingly labeled class receives from the specified predictor variables.

> singleplot(class=2, 'MANIND.M')



Note: Plot partial dependence of MART model on a selected input variable. The bars are ordered in ascending categorical value and the positive / negative length of each corresponding bar represents the value of the partial dependence function for that variable value.

> pairplot(class=2, 'PMT.MTH.D', 'DUPPAY10')

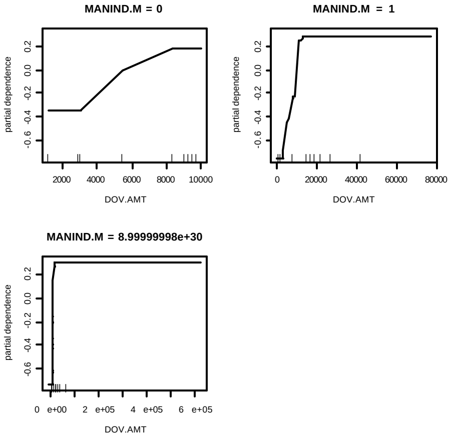


(a)

Note: Plot partial dependence of MART model on a pair of selected input variable. If both variables are categorical, the bars are ordered in ascending categorical value and the positive / negative length of each corresponding bar represents the value of the partial dependence function for that variable value.

Plot *(a)* shows partial dependence of class 2 (CNI) on joint values of the predictor variables "DUPPAY10" and "PMT.MTH.D", i.e., the partial dependence of the model for class 2 on "PMT.MTH.D", conditioned on the variable "DUPPAY10".

> pairplot(class=2, 'DOV.AMT', 'MANIND.M')

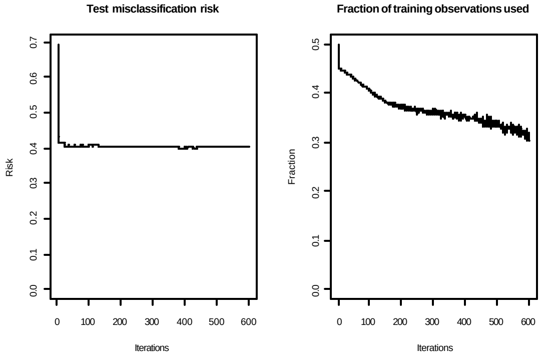


(b)

Note: Plot *(b)* shows fraud class-2 conditional dependence on predictor variable DOVAMT (Numerical), conditioned on each value MANIND.M (Categorical) assumes, including the missing value category.

The hash marks at the base of each plot identify the deciles of the data distribution of the corresponding variables.

> progress()



Note: Monitor progress of MART modeling. The first plot shows the test sample misclassification risk. The second one displays the fraction of training observations used at each iteration.

```
> table(oak.cni.y, martpred(oak.cni.x,probs=F))
```

```
oak.cni.y   1  2   3 4
        1 173  4   4 1
        2  38  0   0 0
        4  56 69 190 0
```

Note: Cross-validation table for true CNI values vs present model predicted CNI's.

Other commands inside R

The following command *dimnames(pred)* gives the predictors names present in the *pred* matrix:

```
> dimnames(oak.kbcni.x)
```

```
[1] "DOV.AMT"  "DISC.AMT" "TRANS.NU" "INV.AWAR" "INV.RECV" "CHK.AWAR"
 [7] "INV.RE1"  "CHK.INV"  "CHK.IN1"  "INTEREST" "MILPAY"   "DBOF"
[13] "BRAC"     "OTHERX"   "UNUSUAL"  "ALLX"     "Y1.PRIOR" "Y1.CUR"
[19] "Y2.PRIOR" "Y2.CUR.1" "Y2.CUR.2" "Y3.PLUS"  "ALL.OTHE" "ENHANCE"
[25] "STE"      "POBOX"    "INV.PAYE" "INV.CNT"  "DOVAMT.2" "DOVAMT.1"
[31] "AVG.5K"   "PAYEE.4"  "MULTI.PA" "MULTI.AD" "INV.SEQ"  "PMT.FREQ"
[37] "PMT.FR1"  "TINS"     "MULTI.TI" "MULTI.2"  "MULTI.3"  "MULTI.4"
[43] "MULTI.EF" "DISCOUNT" "FEW.PYMT" "MISC.OBL" "DUPPAY10" "DUPPAY11"
[49] "NOT.DFAR" "NUMADR.K" "NUMADREE" "NUM.EE.K" "MANIND.A" "MANIND.M"
[55] "PMT.MT.D" "PMT.MT.E" "PMT.TP.C" "PMT.TP.F" "PMT.TP.P"
```

Exporting MART results from R

The following commands illustrate a useful way to export MART results in a txt format, easily readable from Microsoft Excel.

```
> write.table(martpred(nflk.caps, probs =F), 'nflk_cni8_2.txt')
```

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B REFERENCE OF VARIABLES USED IN ANALYSIS

The following table constitutes a reference of (59) variables used in Chapter IV analysis.

Reference Name	Variable Name	Type	Description
DOV.AMT	DOV_AMT	Numeric	Disbursing Office Voucher Amount
DISC.AMT	DISC_AMT	Numeric	Discount Amount
TRANS.NU	TRANS_NUM	Categorical	Number of transactions associated with a single payment
INV.AWAR	INV_AWARD_DT	Numeric	Number of days between invoice date and contract award date
INV.RECV	INV_RECV_AWARD_DT	Numeric	Number of days between invoice received date and award date
CHK.AWAR	CHK_AWARD_DT	Numeric	Number of days between check date and award date
INV.RE1	INV_RECV_INV_DT	Numeric	Number of days between invoice received date and invoice date
CHK.INV	CHK_INV_DT	Numeric	Number of days between check date and invoice date
CHK.IN1	CHK_INV_RECV_DT	Numeric	Number of days between the check date and invoice received date
MILPAY	MILPAY	Categorical	Military Pay Appropriation
DBOF	DBOF	Categorical	DBOF Appropriation
OTHERX	OTHERX	Categorical	X Year Appropriation other than BRAC, DBOF, UNUSUAL
UNUSUAL	UNUSUAL	Categorical	Appropriation = 5188, 5189, 6875, 3880, 3875 or 8164
ALLX	ALLX	Categorical	All X year appropriations
Y1.PRIOR	Y1_PRIOR	Categorical	1 year Expired Appropriation
Y1.CUR	Y1_CUR	Categorical	1 Year current appropriation
Y2.CUR1	Y2_CUR_1ST	Categorical	2 Year Current Appropriation Paid 1st Year
Y2.CUR2	Y2_CUR_2ND	Categorical	2 Year Current Appropriation Paid 2nd Year
Y3.PLUS	Y3_PLUS	Categorical	3 or more year appropriation
ALL.OTHE	ALL_OTHER	Categorical	None of the above appropriations starting with MILPAY
ENHANCE	ENHANCE_PAYEE	Categorical	Flag when Payee found in Remit_L1 field
STE	STE	Categorical	Pymt made to suite address
POBOX	POBOX	Categorical	Payments to POBOX

Reference Name	Variable Name	Type	Description
INV.PAYE	INV_PAYEE	Categorical	Payee with different invoice number lengths
INV.CNT	INV_CNT	Categorical	Contract with different invoice number lengths
DOVAMT.2	DOVAMT_2K	Categorical	DOV_AMT >= to 2000
DOVAMT.1	DOVAMT_1K	Categorical	DOV_AMT >= to 1000
AVG.5K	AVG_5K	Categorical	Average payment amount to payee is >= 5K
PAYEE.4	PAYEE_4_PYMT	Categorical	4 or more payments to the same payee
MULTI.PA	MULTI_PAYEE	Categorical	Multiple payees to the same address
MULTI.AD	MULTI_ADR	Categorical	Muliple address to the same payee
INV.SEQ	INV_SEQ	Categorical	Invoices out of sequence to the same payee
PMT.FREQ	PMT_FREQ_HI	Categorical	Regular payments over a period of time
PMT.FR1	PMT_FREQ_LO	Categorical	Payments occuring in a short period to time
TINS	TINS	Categorical	Tax identification number is present in record
MULTI.TI	MULTI_TINS	Categorical	Multiple TINS for a Payee
MULTI.2	MULTI_PAYEE_K	Categorical	Multiple Payees to the same contract
MULTI.3	MULTI_ADDR_K	Categorical	Multiple Addresses to the same contract
MULTI.4	MULTI_TINS_K	Categorical	Multiple TINS to the same contract
FEW.PYMT	FEW_PYMT	Categorical	Flag companies that have <200 payments in a year
MISC.OBL	MISC_OBLIG	Categorical	Flag that looks for MORD or MOD in the PIIN
DUPPAY10	DUPPAY102	Categorical	Duplicate Payment Indicator 102 - Logic: Same PIIN, Same SPIIN, Same Inv#, DOV Amt >=2000
DUPPAY11	DUPPAY110	Categorical	Duplicate Payment Indicator 110 - Same INV#, Same DOV Amt, DOV Amt >= 2000
NOT.DFAR	NOT.DFAR	Categorical	PIIN/Del Ord does not comform to the DFAR
NUMADR.K	NUMADR_K	Categorical	Number of addresses (ADR_L1+CITY) to an individual contract (PIIN+DO).
NUMADREE	NUMADREE	Categorical	Number of addresses (ADR_L1+CITY) to a whole payee.
NUM.EE.K	NUM_EE_K	Categorical	Number of whole payees to an individual contract (PIIN+DO).
MANIND.A	MAN_IND	Categorical	Manual Indicator {0}
MANIND.M	MAN_IND	Categorical	Manual Indicator {1}

Reference Name	Variable Name	Type	Description
PMT.MT.D	PMT_METH	Categorical	Payment Method {D}
PMT.MT.E	PMT_METH	Categorical	Payment Method {E}
PMT.TP.C	PMT_TYPE	Categorical	Payment Type {C}
PMT.TP.F	PMT_TYPE	Categorical	Payment Type {F}
PMT.TP.P	PMT_TYPE	Categorical	Payment Type {P}
PMT.PR.F	PMT_PROV	Categorical	Payment Provision {F}
PMT.PR.P	PMT_PROV	Categorical	Payment Provision {P}
PMT.PR.R	PMT_PROV	Categorical	Payment Provision {R}
PPA.XM.C	PPA_XMPT	Categorical	Prompt Payment Act Exempt {C}
PPA.XM.E	PPA_XMPT	Categorical	Prompt Payment Act Exempt {E}

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX C FRAUD PROPORTION TEST FOR DIFFERENT SITES

The following table summarizes the selected models and final prediction results for each site studied in the analysis of fraud proportion.

Site	*Model* *MART(M, F)*	*Test Misclassification rate*	*Prediction over Test sets*
Dayton	*MART(138, 4)*	*0.003*	*table(martpred(v.dyt))* 0 1 3 17509 2 49 *table(yv.kb,martpred(v.kb))* 1 2 3 4 1 72 1 0 0 2 0 11 0 0 3 0 0 10 0 4 0 0 0 22
Oakland	*MART(161, 3)*	*0.000*	*table(martpred(v.ok))* 0 1 3 14496 3 1 *table(yv.kb,martpred(v.kb))* 1 2 3 4 1 72 1 0 0 2 0 11 0 0 3 0 0 10 0 4 0 0 0 22
San Antonio [CAPS]	*MART(178, 5)*	*0.000*	*table(martpred(v.sac))* 0 4 17868 2 *table(yv.kb,martpred(v.kb))* 1 2 3 4 1 72 1 0 0 2 0 11 0 0 3 0 0 10 0 4 0 0 0 22

Site	*Model* *MART(M, F)*	*Test Misclassification rate*	*Prediction over Test sets*

San Antonio [IAPS]	*MART(118, 4)*	*0.000*	*table(martpred(v.sai))*
			`    0   1   3 4`
			` 19843 19 10 1`
			table(yv.kb,martpred(v.kb.num))
			`   1   2   3   4`
			`1 72   1   0   0`
			`2  0  11   0   0`
			`3  0   0  10   0`
			`4  0   0   0  22`

San Diego	*MART(140, 4)*	*0.001*	*table(martpred(v.sd))*
			`    0 4`
			` 16900 2`
			table(yv.kb,martpred(v.kb))
			`  0  1   2   3   4`
			`1 0 73   0   0   0`
			`2 0  0  11   0   0`
			`3 0  0   0  10   0`
			`4 2  0   0   0  20`

Pensacola	*MART(145, 4)*	*0.000*	*table(martpred(v.psc))*
			`    0 4`
			` 28397 3`
			table(yv.kb,martpred(v.kb))
			`   1   2   3   4`
			`1 73   0   0   0`
			`2  0  11   0   0`
			`3  0   0  10   0`
			`4  0   0   0  22`

APPENDIX D MISSING VALUES' UNEQUAL VARIANCE t TEST

		DOV.AMT	DISC.AMT	TRANS.NU	INV.AWAR	INV.RECV	CHK.AWAR	INV.RE1	CHK.INV.	CHK.IN1
MAN.IND	t	-4.4	3.3	-.9	10.9	10.1	9.1	-3.6	-6.9	-6.4
	df	398.0	125.0	391.5	194.4	197.2	215.4	439.9	435.9	414.9
	P(2-tail)	.000	.001	.355	.000	.000	.000	.000	.000	.000
	# Present	126	126	126	126	126	126	126	126	126
	# Missing	316	316	316	316	316	316	316	316	316
	Mean(Present)	17934.8	34.8	1.01	550.8	567.8	584.4	17.1	33.6	16.5
	Mean(Missing)	34388.7	.0000	1.02	183.5	227.8	279.2	44.3	95.7	51.3
PMT.METH	t	-6.9	3.2	2.3	13.3	15.7	16.2	7.7	8.5	3.5
	df	2.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0
	P(2-tail)	.020	.001	.019	.000	.000	.000	.000	.000	.001
	# Present	439	439	439	439	439	439	439	439	439
	# Missing	3	3	3	3	3	3	3	3	3
	Mean(Present)	26480.9	9.9	1.02	289.7	326.5	367.9	36.8	78.3	41.5
	Mean(Missing)	500490.3	.000	1.00	76.00	76.0	106.0	.000	30.0	30.0
PMT.TYPE	t	-1.8	3.2	2.3	2.9	3.1	3.5	2.3	5.5	3.8
	df	6.0	434.0	434.0	7.2	7.1	7.1	12.1	48.0	19.2
	P(2-tail)	.114	.001	.019	.023	.017	.010	.043	.000	.001
	# Present	435	435	435	435	435	435	435	435	435
	# Missing	7	7	7	7	7	7	7	7	7
	Mean(Present)	26675.1	10.1	1.0	290.7	327.6	369.3	36.9	78.6	41.7
	Mean(Missing)	217561.4	.000	1.0	133.9	150.9	173.6	17.0	39.7	22.7
PMT.PROV	t	-1.8	3.0	2.4	6.3	6.2	5.4	-1.0	-3.1	-4.0
	df	438.7	174.1	171.0	287.4	297.1	333.8	434.2	430.1	419.1
	P(2-tail)	.066	.003	.019	.000	.000	.000	.342	.002	.000
	# Present	172	172	172	172	172	172	172	172	172
	# Missing	270	270	270	270	270	270	270	270	270
	Mean(Present)	24323.6	24.32	1.0	417.1	448.4	474.6	31.3	57.5	26.2
	Mean(Missing)	33122.1	.7381	1.0	206.2	246.1	297.1	39.9	90.9	51.1

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E CNI’S IMPORTANT PREDICTORS

Figure E.1, to Figure E.3 show the set of important predictors for classifying Serious CNI’s (CNI1), CNI’s (CNI2) and noCNI (CNI4).

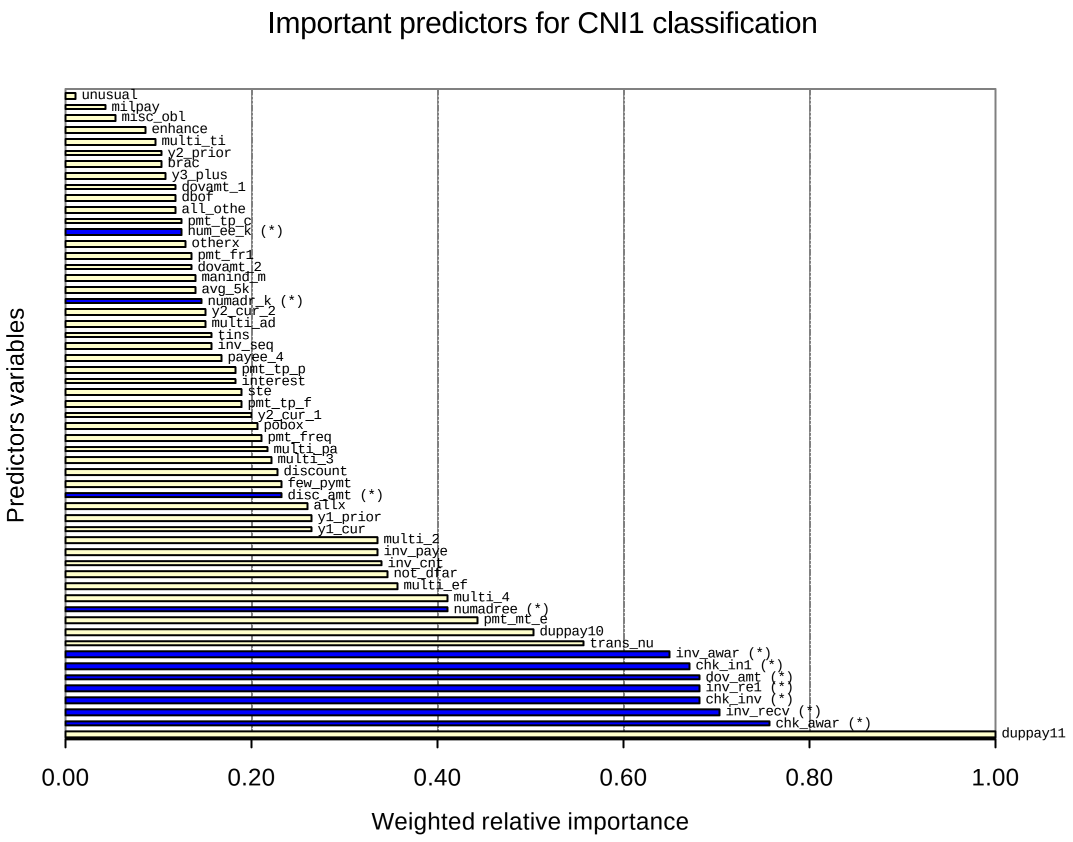


Figure E.1 Variable importance for classifying Serious CNI’s

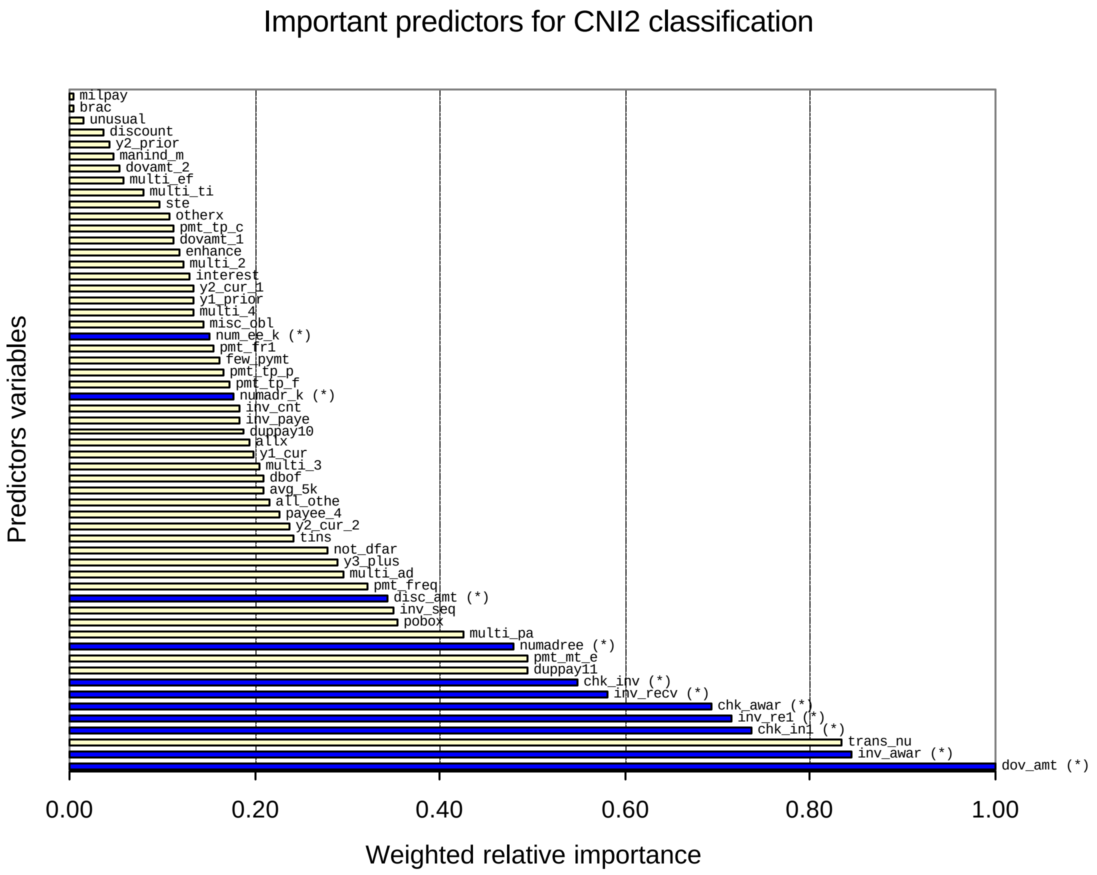


Figure E.2 Variable importance for classifying CNI’s

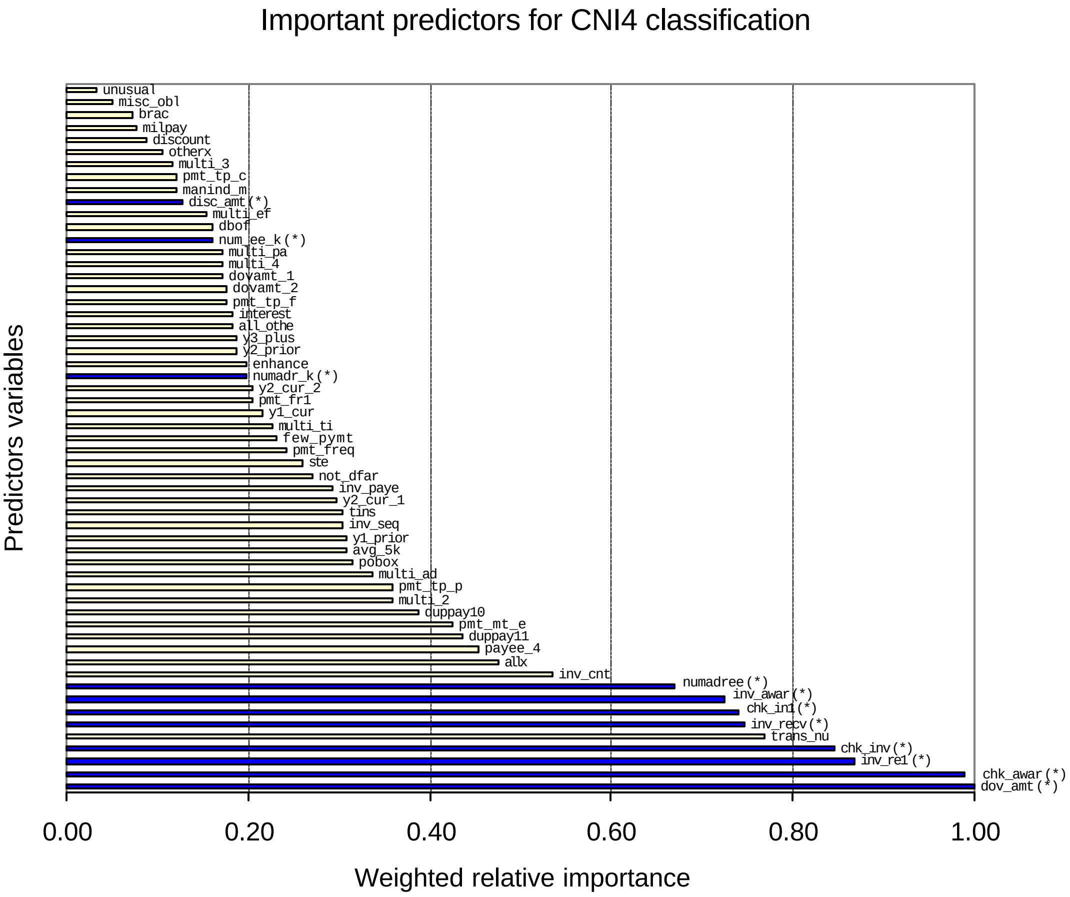


Figure E.3 Variable importance for classifying no-CNI's

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] Breiman, Leo, *Statistical Modeling: The Two Cultures*, Statistical Science, 2001, vol.16, Nr.3, 199-231.

[2] Breiman, Leo, Friedman, J. H., Olshen, R., Stone, C., *Classification and Regression Trees*, Wadsworth, 1983.

[3] Buja, Andreas, Yung-Seoplec, *Data Mining Criteria for Tree-based Regression and Classification*, 2001.

[4] Caruana, Rich, Freitag, Dayne, *Greedy Attribute Selection*, in W. Cohen and H. Hirsh editors, Machine Learning: Proceedings of the Eleventh International Conference, Morgan Kaufman, 1994.

[5] Drucker, Harris, Cortes, Corinna, *Boosting Decision Trees*, Neural Information Processing 8, Morgan Kaufmann, NIPS-8 1996, eds. D.D. Touretsky, MC. Mozer and ME. Hasselmo with C. Cones, MIT Press, pp470-485.

[6] Fujikawa, Yoshikazu, *Efficient Algorithms for Dealing with Missing values in Knowledge Discovery*, Master's Thesis, School of Knowledge Science, Japan Advanced Institute of Science and Technology, February 13, 2001.

[7] Friedman, H. Jerome, *Stochastic Gradient Boosting*, March 26, 1999.

[8] Friedman, H. Jerome, *Tutorial: Getting Started with MART in R*, Stanford University, April 2002.

[9] Friedman, H. Jerome, *Greedy function approximation: A Gradient Boosting machine*, April 19, 2001.

[10] Hand, David, Mannila, Heikki, Smyth, Padhraic, *Principles of Data Mining*, MIT Press, 2001

[11] Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlog, New York, 2001.

[12] Jenkins, Donald, *Evaluation of fraud detection data-mining used in the auditing process of the Defense Finance and Accounting,* Master's Thesis, Operations Research Department, Naval Postgraduate School, Jun 2002.

[13] John, George H., Kohavi, Ron, Pfleger Karl, *Irrelevant Features and the Subset Selection Problem*, Machine Learning: Proceedings of the Eleventh Int'l Conference, Morgan Kaufmann Publishers, San Francisco, CA, pp121-129.

[14] Little, Roderick J. A., Rubin, Donald B., *Statistical analysis with missing data*, Wiley series in probability and mathematical statistics, 1987.

[15] Liu, W. Z., White, A. P., Thompson and Bramer, M. A., *Techniques for Dealing with Missing Values in Classification*, in Second Int'l Symposium on Intelligent Data Analysis, London, 1997.

[16] Mannila, Heikki, *Theoretical Frameworks for Data Mining,* SIGKDD Explorations, ACM SIGKDD, 2000.

[17] Mannila, Heikki, *Data mining: machine learning, statistics and databases,* Eight International Conference on Scientific and Statistical Database Management, Stockholm, June 18-20, 1996.

[18] Quinlan, J. R., *Unknown attribute values in induction*, in Proceedings of the Sixth Int'l Workshop on Machine Learning, Morgan Kaufmann, Los Altos, USA, 1989.

[19] Ragel, Arnaud and Cremilleux, Bruno, *MVC - A Preprocessing Method to Deal With Missing Values*, Knowledge-Based Systems Journal (to appear).

[20] Ripley, B. D., *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press, 1996.

[21] Schapire, Robert E., *A Brief Introduction to Boosting*, Proceedings of the Sixteenth Int'l Joint Conference on Artificial Intelligence, 1999.

[22] Shawn R. Jones-Oxendine, *An Analysis of DOD Fraudulent Vendor Payments*, Master's Thesis, Systems Management Department, Naval Postgraduate School, 1999.

[23] Defense Finance and Accounting Service, *Improper Payments/ Data Mining – Project Support*, Final Report.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft Belvoir, VA 22060-6218

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, CA 93943-5101

3. Director, Studies and Analysis Division
   MCCDC Code C45
   Quantico, VA 22134-5130

4. Direcção de Análise e Gestão de Informação
   Marinha – Portugal

5. Direcção do Serviço de Formação
   Marinha – Portugal

6. Instituto Hidrográfico
   Marinha – Portugal

7. Instituto Superior Naval de Guerra
   Marinha – Portugal

8. Escola Naval
   Marinha – Portugal

9. Dr. Lyn R. Whitaker
   Operations Research Department
   U.S. Naval Postgraduate School
   Monterey, CA 93943

10. Dr. Samuel E. Buttrey
    Operations Research Department
    U.S. Naval Postgraduate School
    Monterey, CA 93943

11. David Riney
    Defense Finance and Accounting Service,
    Operation Mongoose - Internal Review
    Seaside, CA 93955-6771

12. LCDR António Monteiro
    Portuguese Navy